INVESTIGATING PROKARYOTIC TRANSCRIPTOMES AND THE IMPACT OF CROSSTALK BETWEEN NONCODING RNA AND MESSENGER RNA INTERACTIONS

A thesis submitted in partial fulfilment of the requirements

for the Degree of Doctor of Philosophy in Biotechnology

in the University of Canterbury

by Sinan Uğur Umu

University of Canterbury

2016

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisors Paul P. Gardner and Anthony M. Poole for the support during my PhD studies and for selecting me in the first place. Besides my advisors, I would like to thank to rest of my thesis committee for their time.

I would like to thank all the UC Bioinformatics (current and former) members of the Poole & Gardner lab, my co-authors, my colleagues on the 5th floor, the administrative assistants and the other staff of School of Biological Sciences. I especially would like to acknowledge Nicole Wheeler and Bethany Jose for their proof-readings for this thesis.

I would like to acknowledge my funding institutions Biomolecular Interaction Centre and UC HPC (BlueFern) for my PhD scholarship. I also want to thank the staff of UC HPC for their help and assistance.

Last but not the least, I would like to thank my family and my beloved wife, Özgün Candan Onarman Umu. The past three years were hard without her, but she always encouraged me to continue.

ABSTRACT

Prokaryotes have a complex non-coding RNA (ncRNA) based regulatory system, resembling that of eukaryotes. Recent transcriptomics studies also point out the abundance of highly expressed uncharacterized RNAs in archaea and bacteria. However, despite the recent advances indicating the prevalence of ncRNAs in prokaryotes, it is still unknown to what extent these uncharacterized transcripts are functional. Therefore, we have proposed a phylogeny informed approach to design new RNA sequencing (RNA-seq) experiments, which increases the information harnessed from transcriptome data for ncRNA detection.

Many regulatory ncRNAs engage in RNA-RNA interactions, where RNA molecules bind to form a duplex. Predictions of true targets for an RNA enables a successful functional characterization, these can be estimated by bioinformatics methods. However, the algorithms developed to date are imperfect and it is an open question as to which ones perform well and whether these can be improved upon. Towards this goal we performed a computational benchmark study to find reliable algorithms for RNA-RNA interaction prediction. We found that energy based methods, which include the accessibility of interaction regions, are currently the most accurate.

Many ncRNAs, including housekeeping ncRNA genes, are highly expressed. The abundances of interacting RNA molecules enable RNA-RNA duplex formation. In chapter IV we explore the impact of high abundance RNAs on protein expression due to crosstalk RNA-RNA interactions between mRNAs and ncRNAs. With extensive RNA-RNA interaction predictions we reveal that RNA avoidance is an evolutionarily conserved phenomenon among prokaryotes, which means that core mRNAs have evolved to avoid crosstalk interactions with abundant ncRNAs. Our predictions also reveal that RNA avoidance may influence protein expression. To test this, we investigated the stability of interactions between mRNAs and core ncRNAs. These predictions show that the RNA avoidance influences the final protein abundances.

In conclusion, the primary aims of this study are to investigate the prokaryotic transcriptome for novel ncRNA genes and examine the effects of crosstalk RNA interactions. We present a method to increase information gained from transcriptome in prokaryotes for ncRNA identification. We also present the most comprehensive benchmark of RNA-RNA interaction prediction algorithms to date. Lastly, we introduce and test a 'RNA avoidance hypothesis' that shows the influence of crosstalk RNA interactions on protein expression in bacteria.

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 2: Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling
Chapter 3: A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life
Chapter 4: Natural avoidance of stochastic mRNA:ncRNA interactions can be harnessed to control protein expression levels

Please detail the nature and extent (%) of contribution by the candidate:

Sinan was a major contributor to each of the above manuscripts (either first author, or joint first author).

I estimate his % contributions for each manuscript to have been:

**Ch1: 40%*

**Ch2: 90%*

**Ch3: 75%*

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work

- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name: *Paul Gardner*

Signature: 

Date: *2016-01-29*

CHAPTER I - Introduction

1.1 A brief introduction to RNA biology and ncRNAs

RNAs are not only simple carriers of protein information as in messenger RNAs (mRNAs), but they also have prominent roles in cellular regulation. The existence of ncRNA genes was proposed a long time ago (Jacob & Monod 1961), but the study of ncRNA progressed more slowly than protein coding genes. New discoveries in ncRNA biology have emerged in all domains of life within the last decade (eukaryotes, bacteria and archaea) and remarkably altered molecular biology (Cech & Steitz 2014; Sharp 2009). Since RNAs can provide both catalytic activity needed to sustain primitive life and hold genetic information, they even led to the concept of an RNA world where RNAs were the first primitive life forms (Higgs & Lehman 2015; Gilbert 1986). In short, RNAs are a key ingredient for a wide array of cellular functions (Storz et al. 2011; Vogel 2009; Dieterich & Stadler 2012; Burge et al. 2013; Holmqvist & Vogel 2013; Waters & Storz 2009; Cech & Steitz 2014; Borges & Martienssen 2015).

Both eukaryotes and prokaryotes have many ncRNA genes (Cech & Steitz 2014; Borges & Martienssen 2015; Kung et al. 2013; Storz et al. 2011) and large portion of their genomes are pervasively transcribed into RNA (mostly uncharacterized non-coding transcripts) (Kapranov et al. 2007; Wade & Grainger 2014; Djebali et al. 2012). Although all non-coding regions are not necessarily functional genes, only a small fraction of the genome has coding potential (i.e. protein coding genes) in many eukaryotes (Mattick 2004; Mattick 2009). For instance, the human genome contains ~1.2% protein coding mRNAs, while 76% is transcribed into RNA (and ~80% of non-coding regions are biochemically active) (Dunham et al. 2012; Pennisi 2012; Djebali et al. 2012) and it also has more than 100,000 long non-coding RNA (lncRNA) genes that create more than 160,000 lncRNA transcripts (>200 nt long non-coding transcripts) (Zhao et al. 2016).

Once assumed to be relatively simple, prokaryotic transcriptomes have been shown to have transcriptional complexity similar to that of eukaryotes (Barquist & Vogel 2015; Güell et al. 2009; Güell et al. 2011; Sharma et al. 2010). Advancements in sequencing technology also unveil that non-canonical transcripts (e.g. antisense non-coding transcripts) and transcripts without any protein product are common in prokaryotes, referred as the 'genomic dark matter' (Wade & Grainger 2014; Lindgreen et al. 2014; Croucher & Thomson 2010). Furthermore, ~98% of total RNA molecules in a typical bacterial cell are ncRNAs (Deutscher 2003; Deutscher 2006; Giannoukos et al. 2012). Although these are predominantly ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) (Giannoukos et al. 2012), the other ncRNAs also form a significant fraction of the total RNA outputs in prokaryotes (Lindgreen et al. 2014).

In summary, RNAs are versatile and abundant molecules. In the following sections, the properties and features of RNAs will be described in more detail.

1.1.1 RNA structures and structure prediction

RNA molecules interact with proteins and metabolites or form RNA-RNA interactions to execute their functions (Cech & Steitz 2014; Storz et al. 2011; Waters & Storz 2009; Wassarman & Storz 2000; Carthew & Sontheimer 2009; Borges & Martienssen 2015). These functions require that RNAs exist in a form which enables them to interact with other molecules (e.g. oligonucleotides, metabolites and proteins). These evolutionarily conserved secondary structures are usually an important feature of ncRNAs (Wuchty et al. 1999; Dieterich & Stadler 2012; Eddy 2006). These structures also make comparative analyses possible among ncRNAs (Hoeppner et al. 2012; Barquist et al. 2012; Gardner et al. 2005; Nawrocki et al. 2015). For example, the Rfam database holds a collection of these ncRNA structures, represented by covariance models (CMs) (Eddy & Durbin 1994) and multiple sequence alignments (MSAs) (Gardner et al. 2009; Nawrocki et al. 2015).

CMs are a special case of hidden Markov models (HMM) that provide covariation information of base-pairing interactions (Eddy & Durbin 1994). An HMM probabilistic model can describe a

series of observations (e.g. protein and nucleotide strings) based on a training dataset (Krogh et al. 1994). In computational biology, HMMs can use homology information derived from protein or nucleotide MSAs (i.e. training dataset) to detect regions of primary sequence similarity (Krogh et al. 1994; Wheeler & Eddy 2013; Mistry et al. 2013). In a similar fashion, an RNA CM can capture and score all the base-pairing intramolecular interactions as well as primary sequence information for sequence similarity analysis (Eddy 2006; Eddy & Durbin 1994).

There are many RNA sequences available, but experimentally verified RNA structures are in short supply, which makes bioinformatic predictions an important task and an effective *in silico* alternative. The primary structure of an RNA molecule is a simple oligonucleotide that consists of covalently linked A, U, G and C ribonucleotides. RNAs do not stay as linear molecules (including mRNAs and ncRNAs), instead folding onto themselves using base-pairing interactions (Watson-Crick A-U, G-C and wobble G-U) to form RNA secondary structures (Figure 1.1) (Onoa & Tinoco 2004), which are a scaffold for tertiary structures (Fürtig et al. 2003; Tinoco & Bustamante 1999; Onoa & Tinoco 2004; Gardner & Giegerich 2004). It is easier to model secondary structures than tertiary structures; therefore, various algorithms have been developed to predict RNA secondary structure (Zuker & Sankoff 1984; Mathews 2006; Mathews & Turner 2006; Do et al. 2006; Knudsen & Hein 2003).

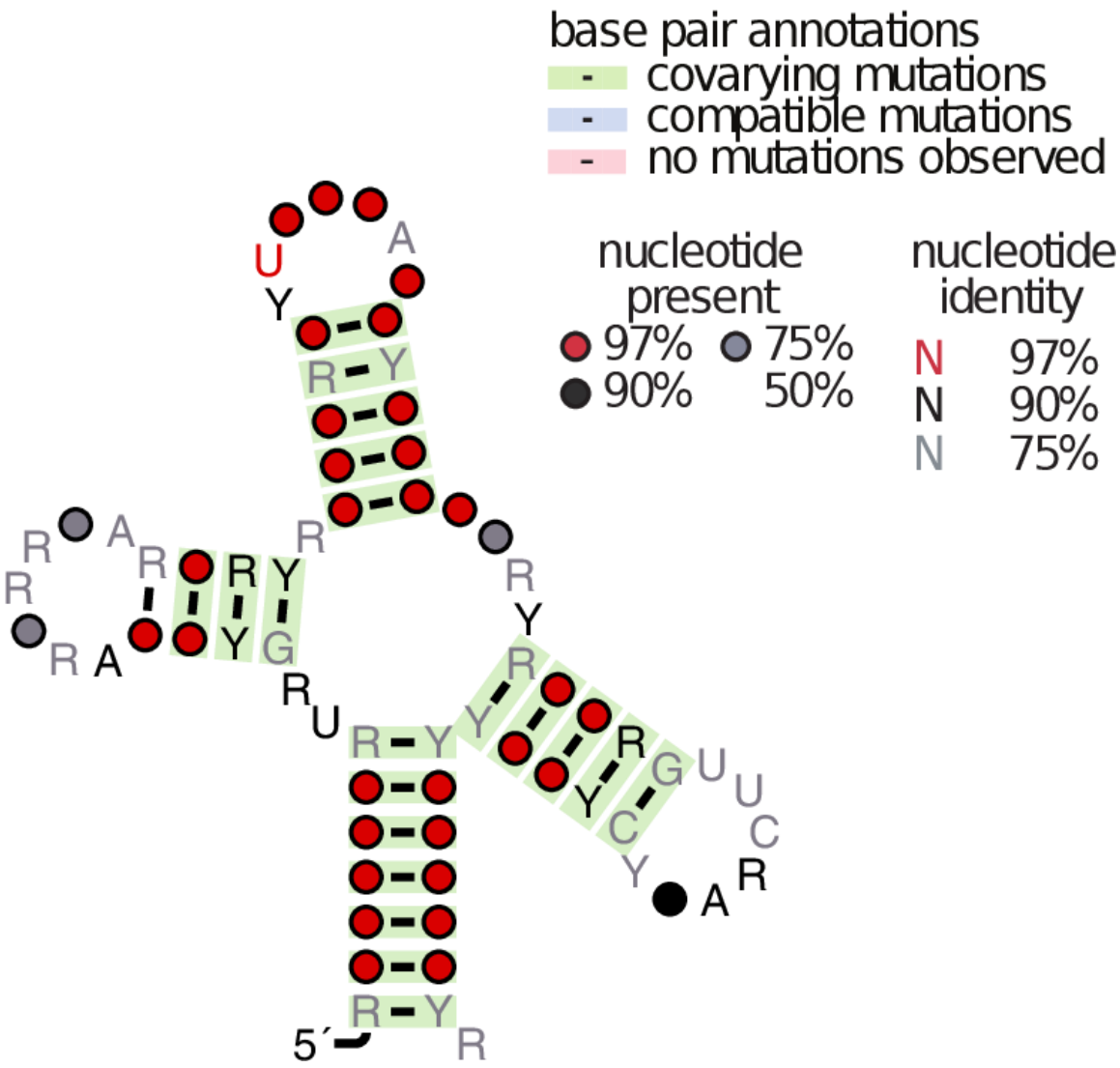

Figure 1.1 The secondary structure of an RNA molecule consists of base pairing interactions of primary structure (i.e. sequence). In this plot the secondary structure of the phenylalanine tRNA of yeast is seen, drawn by the R2R package (Weinberg & Breaker 2011). A conserved secondary structure is an important aspect of most RNA families. In this case the tRNA's cloverleaf shape is visible. The Rfam family RF00005 contains the model of this structure (Nawrocki et al. 2015).

It has been estimated that the number of possible secondary structures for an RNA molecule is greater than 1.8^n, where 'n' is the number of ribonucleotides in the RNA sequence (Zuker & Sankoff 1984; Doshi et al. 2004), which makes RNA secondary structure prediction a difficult problem. Minimum free energy (MFE) methods, which predict the optimal folding energy of

RNA molecules, are widely used to computationally predict RNA secondary structures (Dieterich & Stadler 2012; Zuker & Stiegler 1981; Mathews & Turner 2006; Mathews 2006; Zuker & Sankoff 1984), as they are computationally feasible enough, and allow *ab initio* prediction of structures from a single sequence. MFE methods use a dynamic programming approach and the thermodynamic parameters determined by nearest-neighbor energy model to increase the number of (stacking) base-pairs in a predicted stable RNA structure with the lowest energy (Zuker & Sankoff 1984; Zuker & Stiegler 1981; Doshi et al. 2004). The nearest-neighbor model of the RNA structure (Figure 1.2) assumes that the stability of a base-pair depends on its adjacent base-pairs (Zuker et al. 1999; Xia et al. 1998; Turner et al. 1988), which is a good approximation to predict an RNA secondary structure since the majority of stabilizing interactions are stacking and hydrogen bonding (Turner et al. 1988).

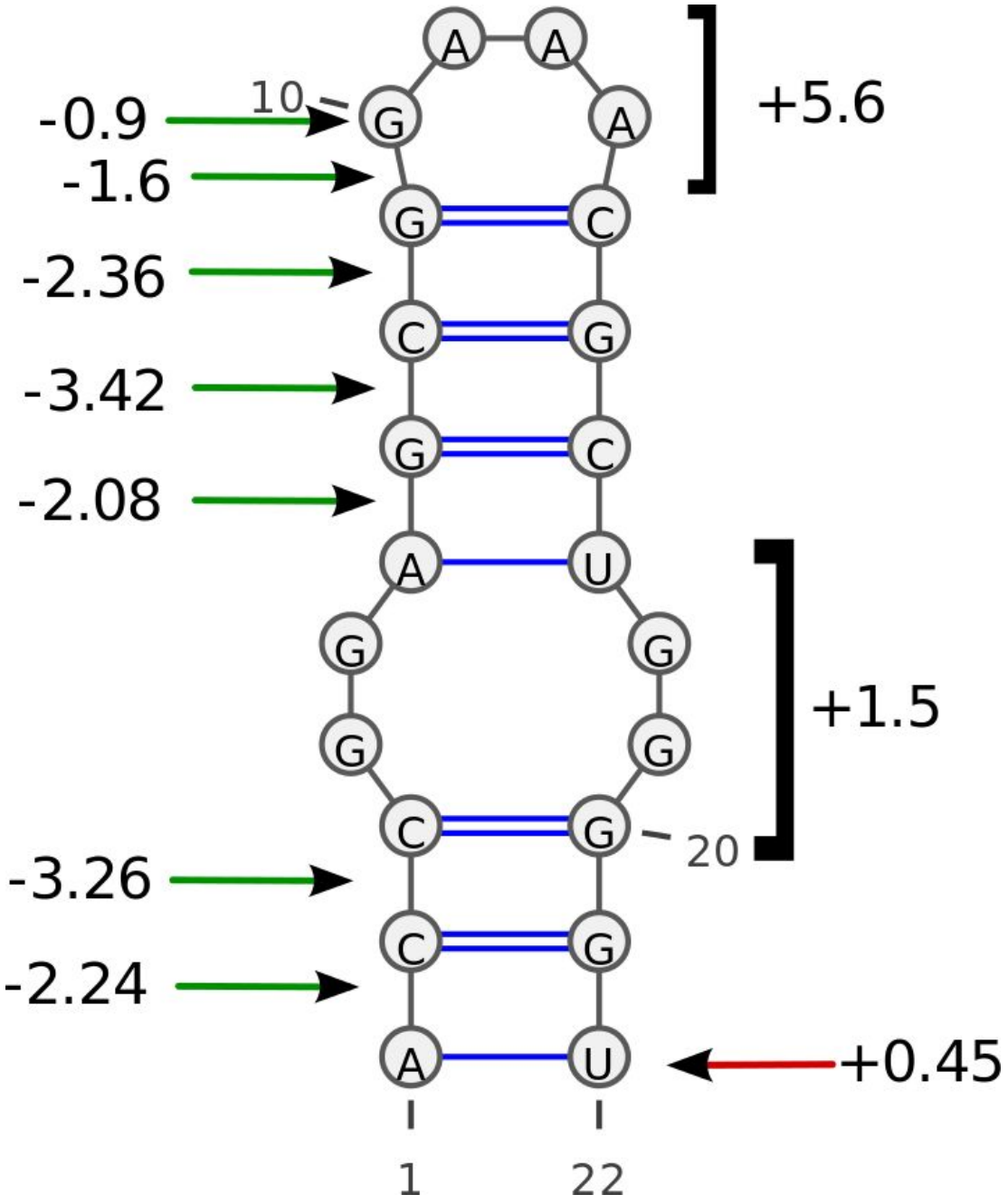

Figure 1.2 An example of nearest neighbor calculation for an RNA molecule. The energy gained from stacks (green arrows) are added for helices. There is a bonus for having GA as the first noncanonical base pair (-0.9 kcal/mol). There is a penalty for helices ending with AU or GU (+0.45 kcal/mol) (red arrow) (Xia et al. 1998). The brackets show the (internal and hairpin) loops which are penalized (+5.46 kcal/mol and +1.5 kcal/mol). The total free energy is the sum of individual terms. We used the VARNA tool (Darty et al. 2009) to redraw this RNA structure,

and the individual stack/loop parameters were taken from (Andronescu et al. 2014), all units are kcal/mol.

RNA-RNA interactions are the formation of RNA duplexes via antisense binding of RNAs and the MFE methods are also popular in interaction predictions (Lorenz et al. 2011; Backofen & Hess 2010; Lai & Meyer 2015; Pain et al. 2015). However, the MFE methods have limited accuracy in both structure (Gardner & Giegerich 2004; Layton & Bundschuh 2005) and interaction predictions (Dieterich & Stadler 2012; Pain et al. 2015; Lai & Meyer 2015). Many algorithms also ignore pseudoknots (Nussinov et al. 1978; Zuker & Stiegler 1981), which are an important RNA motif (Staple & Butcher 2005). They generally have lower accuracies due to limited thermodynamic parameters and approximations (Mathews & Turner 2006; Mathews 2006; Wuchty et al. 1999). Moreover, the lowest free energy is not necessarily the true native structural energy of an RNA, and it might have multiple conformations with different free energies (Mathews & Turner 2006; Mathews 2006). Cellular dynamics (i.e. interactions with other molecules) and ion concentrations may also influence the form of an RNA inside a cell (Onoa & Tinoco 2004), which is hard to model. For example, *in vitro* RNA structures are more visible without the presence of other cellular components (Ramos & Laederach 2014; Rouskin et al. 2013), and Mg^{2+} increases RNA stability (Deutscher 2006; Alemán et al. 2008).

Because of the limitations, MFE methods are sometimes backed by permutation tests, which compare the energy of native structures and interactions with randomly shuffled sequences (negative controls) to determine how statistically significant the predicted structure or interaction is (Workman & Krogh 1999; Gruber et al. 2007; Park et al. 2013; Clote et al. 2005).

Besides the MFE methods, there are comparative and probabilistic methods available for RNA secondary structure predictions, which require prior information (i.e. homologous sequences and MSAs) for prediction (Gardner & Giegerich 2004; Bernhart et al. 2008; Knudsen & Hein 2003; Hofacker et al. 2002; Do et al. 2006). Comparative methods score evolutionary conservation to increase accuracy of structure prediction and integrate it with the structure thermodynamics

(Hofacker et al. 2002; Bernhart et al. 2008). Probabilistic methods use probabilistic models (e.g. stochastic context-free grammars etc.) to parse a MSA (i.e. training dataset) and determine the folding parameters, and then predict a consensus secondary structure based on these (Do et al. 2006; Knudsen & Hein 2003; Knudsen & Hein 1999).

Even though only ncRNAs are considered to have secondary structures, mRNAs (like all RNAs) can form secondary structures (Park et al. 2013; Chamary & Hurst 2005). Stability of mRNA structures (measured by the MFE methods) was found to correlate with gene expression levels and also influences the rate of evolution (Park et al. 2013; Zur & Tuller 2012; Plotkin & Kudla 2011; Kudla et al. 2009). This suggests that mRNA secondary structures may be subject to selection. Furthermore, in the model organism *Saccharomyces cerevisiae,* translational efficiency and final protein abundance depends significantly on mRNA folding and different mRNA regions are under selection for their secondary structures (Zur & Tuller 2012). There is also a universal trend in both prokaryotes and eukaryotes to decrease mRNA stability (i.e. higher MFE) near the translation-initiation site, which was proposed to increase accessibility of the ribosome binding site (RBS) (Gu et al. 2010).

In summary, RNA structure is an important aspect of all RNAs, including ncRNAs and mRNAs. They define how an RNA behaves, so it is vital to predict an accurate secondary structure to infer a correct function and behaviour of an RNA. Yet, bioinformatics methods can produce incorrect predictions for both RNA structure and RNA interaction predictions. Moreover, RNA structures are important features for RNA-RNA interactions, which are summarized below in more detail.

1.1.2 RNA regulators and RNA-RNA interactions in prokaryotes

RNA regulators are ubiquitously found in bacteria (Cech & Steitz 2014; Breaker 2012; Storz et al. 2011; Vogel 2009; Barquist & Vogel 2015; Barrick & Breaker 2007). RNAs have roles in transcription (Zhang & Ferré-D'Amaré 2015; Wassarman & Storz 2000), translation (Waters & Storz 2009; Zhang & Ferré-D'Amaré 2015; Storz et al. 2011; Breaker 2012; Loh et al. 2009), mRNA stability (Waters & Storz 2009), immune response (Barrangou et al. 2007; Bhaya et al.

2011), quorum-sensing (Shao & Bassler 2012; Hammer & Bassler 2007) and DNA maintenance or silencing (Waters & Storz 2009) via base pairing, protein and nucleic acid binding (Waters & Storz 2009; Gottesman 2004; Vogel 2009; Barquist & Vogel 2015; Breaker 2012; Bhaya et al. 2011; Barrangou et al. 2007). Surprisingly, there are ~100 regulatory small RNAs (sRNAs) within model bacterial species such as *Escherichia coli* (*E. coli)* and *Salmonella enterica* (Vogel & Luisi 2011; Holmqvist & Vogel 2013). Furthermore, the small genome of *Mycoplasma genitalium (M. genitalium)* contain more than 400 antisense small ncRNAs (Chen et al. 2016), although the functions of antisense RNAs are disputed and they might be a result of the high A+T content (Lloréns-Rico et al. 2016).

Bacterial sRNAs are a subset of regulatory ncRNA families that either repress or activate target transcript expression (mostly repress) via imperfect antisense base-pairing interactions (Storz et al. 2011; Waters & Storz 2009; Shao & Bassler 2012; Jørgensen et al. 2013), which are usually mediated by the RNA chaperone *Hfq* protein (Vogel & Luisi 2011; Jørgensen et al. 2013). *Hfq* protein facilitates the base pairing between RNAs (Vogel & Luisi 2011; Gottesman 2004; Holmqvist et al. 2016) and also protects sRNAs from cellular degradation (Vogel & Luisi 2011; Holmqvist et al. 2016). Some of the sRNAs are also responsible for regulation of more than one target transcript (Andrade et al. 2013; Shao & Bassler 2012), regulation of a network of genes (Jørgensen et al. 2013; Faner & Feig 2013) or co-regulation of a gene with other sRNAs (Modi et al. 2011). These RNAs commonly base pair at or near the ribosome binding site (RBS), located in the 5` untranslated region (UTR), of target mRNA that prevents translation by blocking the ribosome (Bouvier et al. 2008). It is also possible for sRNAs to bind at coding regions (Waters & Storz 2009; Bouvier et al. 2008) or more distant locations from the RBS (Shao & Bassler 2012), which may activate target expression (Waters & Storz 2009; Hammer & Bassler 2007).

In bacterial cells, even cis-regulatory RNA elements like riboswitches can sometimes use trans-acting antisense binding mechanisms to decrease expression of a target transcript (Loh et al. 2009). Furthermore, clustered regularly-interspaced short palindromic repeat (CRISPR) RNAs block invasion of foreign genetic material (by oligonucleotide binding) and act like

bacterial immune response (Barrangou et al. 2007; Bhaya et al. 2011). CRISPR provides adaptive resistance against viruses or plasmids via integrating a piece of foreign material into genome as spacer sequences (Rath et al. 2015).

Archaeal species are known to have small nucleolar RNAs (snoRNAs) like eukaryotes, which have roles in rRNA maturation (Kiss 2002; Omer et al. 2000). They also contain riboswitches (cis regulatory RNA structures) (Barrick & Breaker 2007), CRISPR (Barrangou et al. 2007; Bhaya et al. 2011) and archaeal regulatory sRNAs (Babski et al. 2014; Prasse et al. 2013). Recent studies also show that huge repositories of uncharacterized non-coding transcripts are available in archaeal genomes (Lindgreen et al. 2014).

In summary, both bacteria and archaea contain various regulatory and base-pairing ncRNAs which have fundamental roles in prokaryotic cells.

1.1.3 RNA-RNA interactions and ncRNAs in eukaryotes

Eukaryotes have a very complex ncRNA based regulatory system (Cech & Steitz 2014) especially higher eukaryotes (e.g. mammals and plants). Eukaryotic RNA-RNA interactions mostly focus on RNA interference (RNAi) (Mello & Conte 2004), which means regulation of target genes (usually target expression inhibition) by small RNA binding, microRNAs (miRNAs) and small interfering RNAs (siRNAs) (Carthew & Sontheimer 2009; Ambros 2004; Chen 2008; Oğul et al. 2011; Borges & Martienssen 2015). Both miRNAs (around 20 nucleotides) and siRNAs (around 30 nucleotides) manifest their functions through antisense base-pairing in plants and animals (Carthew & Sontheimer 2009; Ameres & Zamore 2013). In animal genomes miRNAs prefer perfect complementarity in the seed region (from 2nd to 7th nucleotide) and lower complementarity than their plant counterparts (Ambros 2004; Carthew & Sontheimer 2009; Axtell et al. 2011; Tat et al. 2016; Ameres & Zamore 2013). The target binding region generally lies in the 3`UTR of mRNAs (Ambros 2004). It is possible for a miRNA to target more than one region in animals, which is known to increase the efficiency of target gene downregulation (Millar & Waterhouse 2005). In plants, highly complementary target regions of

miRNAs (mostly a single miRNA) may lie in the coding region as well as UTRs rather than only 3`UTRs (Millar & Waterhouse 2005; Axtell et al. 2011; Ameres & Zamore 2013), which is also true for endogenous plant siRNAs (Carthew & Sontheimer 2009; Addo-Quaye et al. 2008). High complementarity binding usually leads to total degradation of the target mRNA (Ambros 2004; Carthew & Sontheimer 2009).

Piwi-associated RNAs (piRNAs) are also small endogenous eukaryotic RNAs, 24-30 nucleotides long (Klattenhoff & Theurkauf 2008). They regulate transposon activity and have functions in preserving germline genome integrity (Klattenhoff & Theurkauf 2008; Zhang et al. 2015; Gou et al. 2015). Some members of piRNAs use antisense binding to regulate target RNAs (Gou et al. 2015) like miRNAs and siRNA.

Besides regulatory short RNAs, there are various other ncRNAs which utilize RNA-RNA interactions in eukaryotes. For instance, H/ACA and C/D snoRNAs have roles in rRNA and small nuclear RNA (snRNA) maturation (Brown et al. 2001; Kiss 2002; Gardner et al. 2010). Spliceosomal snRNAs form ribonucleoprotein (RNP) complexes with other snRNAs and have roles in RNA splicing (Karijolich & Yu 2010), and they are also targeted by snoRNAs (Darzacq et al. 2002). It also seems that some lncRNAs may engage RNA-RNA interactions (Lee et al. 1999; Kung et al. 2013). This group of ncRNAs are also highly expressed in higher eukaryotes (Zhao et al. 2016).

In summary, both prokaryotes and eukaryotes contain a rich repertoire of ncRNAs which usually interact with other molecules for functionality.

1.1.4 Bioinformatics of RNA-RNA interactions

Many regulatory RNAs utilize RNA-RNA interactions to perform their roles (e.g. small RNAs of eukaryotes and prokaryotes). Yet, like experimentally verified RNA structures, experimentally verified RNA-RNA interactions are in short supply. For example, for the model bacteria *E. coli* and *Salmonella*, there are only a few (around 50 and 20 respectively) verified interaction pairs

available (Wang et al. 2016), and many clades of life are not represented in the available databases (Wang et al. 2016; Chou et al. 2016).

Therefore, different algorithms have been developed to predict RNA-RNA interactions (Pain et al. 2015; Lai & Meyer 2015) as an efficient alternative to wet-lab experiments. The methods of RNA-RNA interaction predictions are mainly divided into three major groups, alignment-like methods (Wenzel et al. 2012; Hodas & Aalberts 2004; Gerlach & Giegerich 2006), MFE methods and comparative (homology) methods (Kery et al. 2014; Wright et al. 2013; John et al. 2004). We can also further divide MFE methods into three different subclasses (Figure 1.3); the approaches neglecting intramolecular structure (Rehmsmeier et al. 2004; Lorenz et al. 2011; Reuter & Mathews 2010), the approaches considering intramolecular base-pairs (internal structure) (Andronescu et al. 2005; Reuter & Mathews 2010) and the approaches measuring accessibility of binding regions (Lorenz et al. 2011; Mückstein et al. 2006). There are also other machine learning algorithms (Yang et al. 2008; Oğul et al. 2011), and probabilistic approaches like RactIP (Kato et al. 2010) which uses the CONTRAfold (Do et al. 2006) model for RNA-RNA interaction prediction.

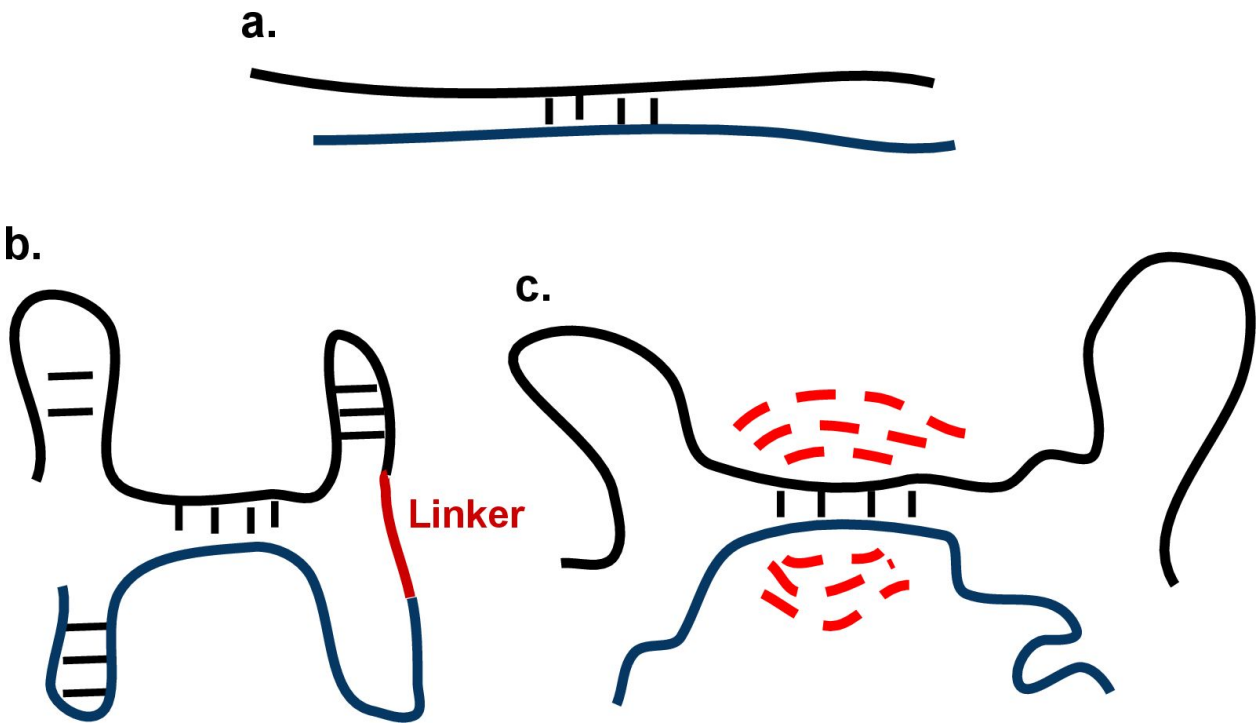

Figure 1.3 (a) The approaches neglecting intramolecular base-pairs only calculate the duplex energy. **(b)** The approaches considering intramolecular base-pairs (i.e. the concatenation approaches) predict a common secondary structure. The two RNA sequences are joined with a linker symbol. **(c)** The accessibility based approaches calculate the energies required to unfold the interaction sites and then combine them with the duplex energy. The red dashed lines represent the unfolded internal base-pairings.

Even though the MFE methods are widely used, they have limited accuracy in both structure (Gardner & Giegerich 2004; Layton & Bundschuh 2005) and interaction predictions (Dieterich & Stadler 2012; Pain et al. 2015; Lai & Meyer 2015). Furthermore, the RNA-RNA interaction prediction algorithms have been developed and benchmarked using a small number of verified interaction pairs, which are only a small fraction of the whole. To overcome this problem, some algorithms focus only a certain type of RNA pairs. For example, algorithms like PLEXY (Kehr et al. 2011) and RNAsnoop (Tafer et al. 2010) predict snoRNA-target interactions; whereas RNAhybrid (Rehmsmeier et al. 2004) and RNAcofold (Hofacker et al. 1994; Bernhart et al. 2006) focus on miRNAs-target interactions.

In summary, a successful prediction of an RNA-RNA interaction is an important bioinformatics problem. Many algorithms have been developed to predict true interacting RNA pairs and most of which are MFE methods. Yet, the algorithms developed have limited accuracy (Pain et al. 2015; Lai & Meyer 2015).

1.2 Characterization, discovery and annotation of ncRNAs

Both prokaryotes and eukaryotes contain ncRNAs with various regulatory functions (Cech & Steitz 2014). Increasing numbers of transcriptome studies have also revealed large regions of (mostly non-coding) uncharacterised transcripts in genomes (Lindgreen et al. 2014; Dunham et al. 2012; Chen et al. 2016; Harrow et al. 2012; Croucher & Thomson 2010; Kapranov et al. 2007), referred to as 'genomic dark matter' (Wade & Grainger 2014; Freyhult et al. 2007; van Bakel et al. 2010; Baboo & Cook 2014).

On the other hand, characterization of functional RNAs (Eddy 2006; Wassarman & Storz 2000; Vogel & Sharma 2005) and illuminating the genomic dark matter (Wade & Grainger 2014; Freyhult et al. 2007; van Bakel et al. 2010; Baboo & Cook 2014) are not easy tasks. Moreover, classical sequence search methods are often not ideal for ncRNA annotation and characterization (Freyhult et al. 2007).

For example, it took decades to understand the role of 6S ncRNA (Brownlee 1971; Hindley 1967) in bacterial cells, which are highly expressed in stationary phase and considered to be a housekeeping gene (Wassarman & Storz 2000). After 30 years of investigation, it was found that 6S regulates RNA polymerase activity and inhibits gene expression (Wassarman & Storz 2000).

In eukaryotes lncRNAs have attracted a lot of attention, since eukaryotes contain large amount of lncRNA regions and transcripts (Zhao et al. 2016; Mattick & Rinn 2015). It is difficult to classify them as protein coding or novel ncRNA genes (Pauli et al. 2015). For example, mouse *Xist* gene, which has a role in X chromosome inactivation (Lee et al. 1999), was misclassified as protein

coding gene containing a translated open reading frame (ORF), but it turned out to be a ncRNA gene (Brockdorff et al. 1992).

One major problem of genome annotation is the protein-centered habits of annotation pipelines (Aziz et al. 2008), which may ignore a range of ncRNAs (Birney et al. 2007). Therefore, many microbial genome sequences in public databases contain limited ncRNA annotations.

In summary, there is speculation on the extent that these non-coding transcripts and pervasive transcription products are functional or are a product of biological noise (van Bakel et al. 2010; Kapranov et al. 2007; Lloréns-Rico et al. 2016; Wade & Grainger 2014; Pauli et al. 2015; Cohen et al. 2016). Thus, successful annotation and characterization of a genomic region (or a transcript) is an important molecular biology problem.

1.2.1 Bioinformatics approaches for discovery of ncRNAs

As with RNA secondary structure and interaction predictions**,** computational tools are an efficient way to discover and annotate ncRNAs. For example, sequence comparison tools like BLAST (Altschul et al. 1990), BLAT (Kent 2002), FASTA (Pearson & Lipman 1988) or similar methods are the gold standard for most of the gene discovery and annotation pipelines (Aziz et al. 2008; Seemann 2014). They are quite successful for protein coding genes, but ncRNAs are known for low primary structure conservation (Freyhult et al. 2007; Eddy & Durbin 1994).

RNA structural features and RNA motifs (Gardner & Eldai 2015; Hofacker et al. 2002; Eddy & Durbin 1994; Eddy 2006) are generally harnessed for functional annotation and/or discovery of ncRNA genes (Nawrocki et al. 2015; Pedersen et al. 2006; Freyhult et al. 2007; Eddy & Durbin 1994; Eddy 2006). RNAs usually have stable secondary structures, which can be detected by using carefully designed negative controls (e.g. dinucleotide preserved shuffled sequences) (Clote et al. 2005; Workman & Krogh 1999). On the other hand, sometimes ncRNA structural energies are not very significant, and hard to separate from background (Rivas & Eddy 2000).

Another problem of ncRNA detection is that most of the ncRNA families are not well conserved, and appear to be evolutionarily young (Lindgreen et al. 2014; Hoeppner et al. 2012) and usually heterogeneous (e.g. size and structure) (Wright et al. 2013). For example, only a few RNA families are conserved among all the domains of life, and those predominantly belong to the protein expression machinery (Hoeppner et al. 2012). Likewise, miRNAs of plants and animals do not share any homologs and are usually restricted to a certain species or taxa (Wilbert & Yeo 2011; Cuperus et al. 2011). Such low conservation of ncRNA genes negatively influences comparative methods and impedes ncRNA discovery. However, comparative methods are an effective way of characterization and annotation if there is prior knowledge (i.e. homologous sequences or sequence alignments) available (Freyhult et al. 2007; Washietl & Hofacker 2004). For instance, the Rfam database collects CM files for each group of homologous ncRNA families (Nawrocki et al. 2015). These can be used by *cmsearch* tool from the Infernal package to annotate ncRNAs in genomes and other sequence-based datasets (Nawrocki et al. 2009; Nawrocki et al. 2015). Furthermore, tools like RNAz (Gruber et al. 2007), QRNA (Rivas & Eddy 2001) and EvoFold (Pedersen et al. 2006) use homology (i.e. conservation of primary and covariation) to discover functional ncRNA genes from alignments. RNAz detects functional ncRNAs from MSAs using structural conservation and thermodynamic stability (Gruber et al. 2007). It predicts consensus MFE structure (Gruber et al. 2007) using RNAalifold (Lorenz et al. 2011; Bernhart et al. 2008), a comparative structure prediction algorithm. QRNA scores pairwise alignments for coding or non-coding potential (Rivas & Eddy 2001). EvoFold takes MSA and phylogenetic tree data, then reports predicted secondary structure with folding potential score (Pedersen et al. 2006). Moreover, tools like RNAcode looks for a coding potential using phylogeny which can separate genes with coding potential (Washietl et al. 2011). A similar tool called PhyloCSF uses comparative genomics to discriminate between a coding region and a non-coding region (Lin et al. 2011).

Protein coding genes contain a set of sequence features like translated ORF signals, hexamer frequency, and codon bias (Rivas & Eddy 2000; Zhang 2002; Andersson & Kurland 1990) which can be exploited by gene finding algorithms (Hyatt et al. 2010; Brettin et al. 2015; Wang et al.

2013) for *ab initio* gene detection. Although, there are not many sequence features available for ncRNA genes (Rivas & Eddy 2001) (except secondary structure and covariation signals), nucleotide distributions may vary between ncRNAs and protein coding regions (Schattner 2002; Klein et al. 2002; Umu et al. 2015). Following that, prokaryotic genomes have various mono-nucleotide distributions (Hurst & Merchant 2001), and ncRNA genes have been identified especially in A+T rich genomes using nucleotide variations (Klein et al. 2002). Moreover, some other sequence features can be used for ncRNA detection such as length, genomic arrangements, terminator/promoter signals and hairpin structures (Vogel & Sharma 2005; Katiyar et al. 2012; Yang et al. 2013), which can be mined by machine learning algorithms (Ogul et al. 2013; Xue et al. 2005; Wu et al. 2011; Kadri et al. 2009).

In summary, low conservation and low primary structure conservation of ncRNA genes impede discovery and annotation of novel ncRNAs. That is why different methods and sequence features are usually utilized for ncRNA discovery.

1.2.2 Experimental methods for ncRNA discovery

The bioinformatics approaches are efficient alternatives, but wet-lab experiments (e.g. genetic screening) are the only way for ultimate validation of mechanism and phenotype of a ncRNA (or any unknown RNA) (Hüttenhofer & Vogel 2006). Experimental methods for a ncRNA identification usually start with cloning by cDNA library preparation or RNA sequencing (replacing tiling-arrays) (Hüttenhofer et al. 2002; Hüttenhofer & Vogel 2006; Yan et al. 2012). Nowadays, low-cost RNA-seq experiments lead to high-throughput detection of novel transcripts for different species and for different stress conditions (Lindgreen et al. 2014; Barquist & Vogel 2015; Croucher & Thomson 2010; Chen et al. 2016; Güell et al. 2009), and differential RNA-seq (dRNA-seq) provides better expression profiles (by discriminating primary transcripts and processed transcripts) that makes ncRNA discovery easier (Sharma & Vogel 2014). For instance, dRNA-seq of *Helicobacter pylori* showed that it contains various novel ncRNAs (e.g. bacterial sRNAs), despite the fact that it does not have *Hfq* chaperone protein (Sharma et al. 2010).

The high-throughput methods (e.g. arrays, RNA-seq, dRNA-seq) often provide hundreds of potentially functional ncRNAs (Barquist & Vogel 2015; Pauli et al. 2015; Chen et al. 2016), and various other high-throughput methods may help to create a shorter list (Yan et al. 2012) such as cross-linking and immunoprecipitation sequencing (CLIP-seq) (Hafner et al. 2010; Riley & Steitz 2013; Holmqvist et al. 2016), ribosome profiling (Ingolia et al. 2009), Tn-seq/TraDIS (van Opijnen et al. 2009; Barquist, Boinett, et al. 2013), RNA structure probing methods (e.g. structure-seq, Mod-seq) (Ding et al. 2014; Underwood et al. 2010; Loughrey et al. 2014; Talkish et al. 2014) and RNA-RNA interaction probing methods (Sharma et al. 2016; Lu et al. 2016; Lu & Chang 2016).

Both in prokaryotes and eukaryotes, ncRNAs usually form RNP complexes or interact with proteins (Hafner et al. 2010; Riley & Steitz 2013). RNA and RNA binding proteins (RBP) are crosslinked to determine RBP binding positions (Hafner et al. 2010; Riley & Steitz 2013) in a CLIP-seq experiment. For example, in bacteria, detection of *Hfq* or *CsrA* binding sites via CLIP-seq allows classification of candidate sRNAs and identification of their possible mRNA targets (Holmqvist et al. 2016; Faner & Feig 2013). Ribosome profiling is the deep sequencing of ribosome protected mRNA fragments (Ingolia et al. 2009). It helps to determine the coding potential of an unknown transcript and can reveal whether a transcript contains a translated ORF or not. (Pauli et al. 2015; Ingolia et al. 2009). Furthermore, ribosome profiling provides translational data which can be used to capture sRNA regulation pathways and candidate targets (Barquist & Vogel 2015). In the techniques like Tn-seq/TraDIS, transposons are inserted into random genomic locations to create mutant strains which are used to determine the regions of functionality on genomes considering various growth conditions (van Opijnen et al. 2009; Barquist, Langridge, et al. 2013; Barquist, Boinett, et al. 2013). Determining the secondary structures of RNAs in a cell by various RNA sequence probing methods help to discover secondary structures of cellular RNAs, which can also help to determine RNA motifs to infer function (Ding et al. 2014; Talkish et al. 2014; Lu & Chang 2016). Likewise, 'LIGation of interacting RNA followed by high-throughput sequencing' (LIGR-seq) (Sharma et al. 2016) and similar methods (Lu et al. 2016; Lu & Chang 2016; Aw et al. 2016) are used to determine

interacting RNAs in cells. These methods help to discover novel RNA-RNA interactions and unveil RNA folding.

Besides high-throughput methods, the potentially functional genes are usually verified by genetic screening methods such as knockouts, overexpression analyses, gene mutant analysis (Ambros 2004; Kuhn et al. 2008; Vogel & Wagner 2007; Addo-Quaye et al. 2008; Yan et al. 2012),

In summary, a combination of bioinformatics and experimental methods help to create a shortlist of potentially functional RNAs for further experimental validation. Since there are not a surplus of any resources in research, the complementary bioinformatics methods of RNA biology (e.g. interaction and structure predictions, ncRNA gene models for annotations etc.) usually provide the researchers with efficient *in silico* alternatives.

1.3 Conclusion

RNAs are versatile and abundant molecules of living cells from all domains of life. It is clear that non-coding transcript discovery, characterization, functional annotation are important topics of molecular and computational biology. Here we have summarized biology of RNA and important features of RNAs.

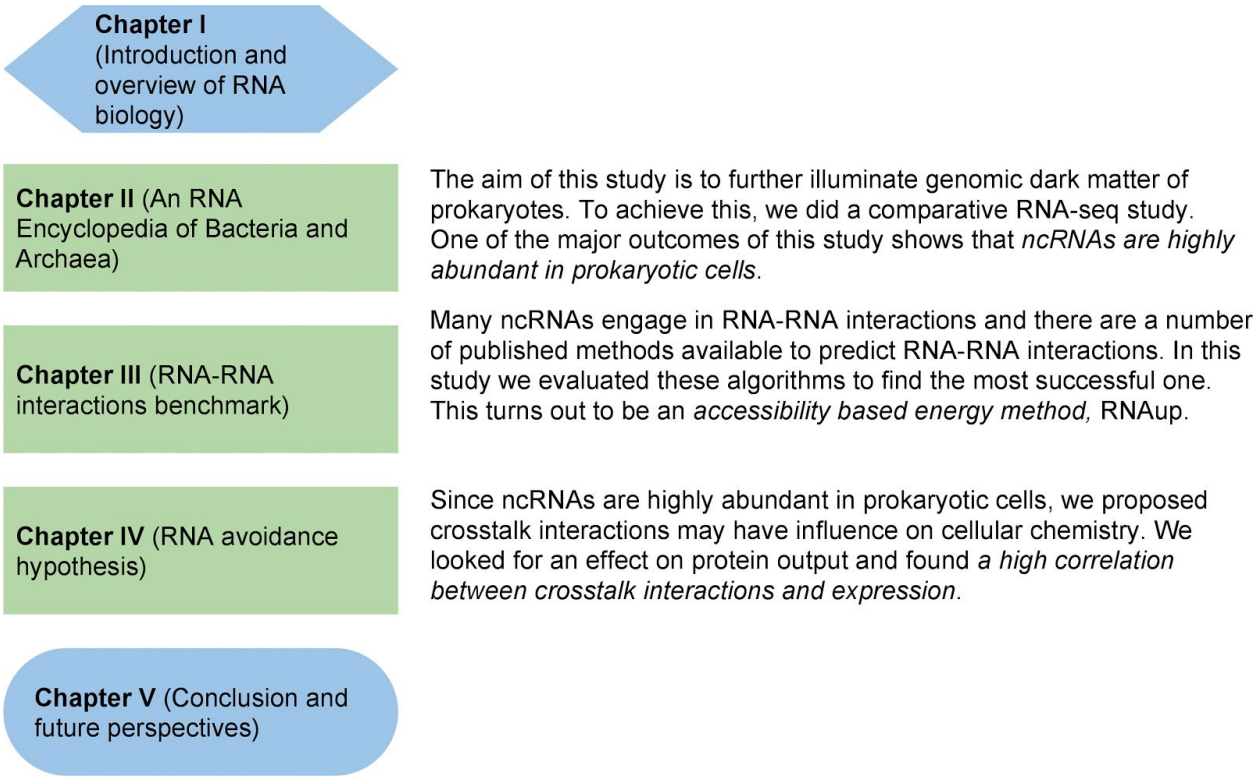

Figure 1.4 An overview of this thesis.

The primary aims of this thesis are to investigate prokaryotic transcriptomes for novel ncRNA genes and examine the influence of crosstalk RNA-RNA interactions in bacteria (Figure 1.4.). To achieve this, first we reexamined available RNA-seq data of archaea and bacteria (Chapter II). This enabled us to propose a new experimental setting to design transcriptomics studies for prokaryotes, which increases the information gained from comparative RNA-seq for ncRNA discovery and annotation. Next, we inspected the current RNA-RNA interaction prediction algorithms (Chapter III). We determined the most successful algorithm with a comprehensive RNA target prediction benchmarking study, which is an important contribution to current literature due to its sample size and algorithm diversity. Finally, we introduce the RNA avoidance hypothesis that shows mRNAs have evolved to avoid crosstalk RNA interactions with core ncRNAs. We tested the effect of crosstalk interactions on protein abundance and compare its effect size with other well-known major factors; mRNA secondary structure and codon bias (Chapter IV). We found that if an mRNA does not readily bind to abundant ncRNAs, it exhibits

higher levels of its translated protein product. Our RNA avoidance model complements the current bacterial protein expression models by explaining the variance among protein-per-mRNA ratio that is not accounted for by the other two factors. Therefore, both our computational predictions and associated experimental verifications prove that RNA avoidance is an important factor in prokaryotic translational efficiency. We also report for the first time evidence that the RNA avoidance signals in mRNAs are evolutionarily conserved across prokaryotic genomes. Lastly, we present a summary and future perspectives of this thesis (Chapter V).

REFERENCES

Addo-Quaye, C. et al., 2008. Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Current biology: CB*, 18(10), pp.758–762.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Ambros, V., 2004. The functions of animal microRNAs. *Nature*, 431(7006), pp.350–355.

Ameres, S.L. & Zamore, P.D., 2013. Diversifying microRNA sequence and function. *Nature reviews. Molecular cell biology*, 14(8), pp.475–488.

Andersson, S.G. & Kurland, C.G., 1990. Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2), pp.198–210.

Andrade, J.M., Pobre, V. & Arraiano, C.M., 2013. Small RNA modules confer different stabilities and interact differently with multiple targets P. Sumby, ed. *PloS one*, 8(1), p.e52866.

Andronescu, M. et al., 2014. The determination of RNA folding nearest neighbor parameters. *Methods in molecular biology* , 1097, pp.45–70.

Andronescu, M., Zhang, Z.C. & Condon, A., 2005. Secondary structure prediction of interacting RNA molecules. *Journal of molecular biology*, 345(5), pp.987–1001.

Aw, J.G.A. et al., 2016. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular cell*, 62(4), pp.603–617.

Axtell, M.J., Westholm, J.O. & Lai, E.C., 2011. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome biology*, 12(4), p.221.

Aziz, R.K. et al., 2008. The RAST Server: rapid annotations using subsystems technology. *BMC*

genomics, 9, p.75.

Baboo, S. & Cook, P.R., 2014. "Dark matter" worlds of unstable RNA and protein. *Nucleus* , 5(4), pp.281–286.

Babski, J. et al., 2014. Small regulatory RNAs in Archaea. *RNA biology*, 11(5), pp.484–493.

Backofen, R. & Hess, W.R., 2010. Computational prediction of sRNAs and their targets in bacteria. *RNA biology*, 7(1), pp.33–42.

van Bakel, H. et al., 2010. Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLoS biology*, 8(5), p.e1000371.

Barquist, L., Langridge, G.C., et al., 2013. A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. *Nucleic acids research*, 41(8), pp.4549–4564.

Barquist, L., Boinett, C.J. & Cain, A.K., 2013. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA biology*, 10(7), pp.1161–1169.

Barquist, L., Burge, S.W. & Gardner, P.P., 2012. Building non-coding RNA families. *arxiv.org*, p.24.

Barquist, L. & Vogel, J., 2015. Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annual review of genetics*, 49, pp.367–394.

Barrangou, R. et al., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), pp.1709–1712.

Barrick, J.E. & Breaker, R.R., 2007. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11), p.R239.

Bernhart, S.H. et al., 2006. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for molecular biology: AMB*, 1(1), p.3.

Bernhart, S.H. et al., 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics*, 9, p.474.

Bhaya, D., Davison, M. & Barrangou, R., 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, pp.273–297.

Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Borges, F. & Martienssen, R.A., 2015. The expanding world of small RNAs in plants. *Nature reviews. Molecular cell biology*, 16(12), pp.727–741.

Bouvier, M. et al., 2008. Small RNA binding to 5' mRNA coding region inhibits translational initiation. *Molecular cell*, 32(6), pp.827–837.

Breaker, R.R., 2012. Riboswitches and the RNA world. *Cold Spring Harbor perspectives in biology*, 4(2). Available at: http://dx.doi.org/10.1101/cshperspect.a003566.

Brettin, T. et al., 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, 5, p.8365.

Brockdorff, N. et al., 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71(3), pp.515–526.

Brown, J.W. et al., 2001. Multiple snoRNA gene clusters from Arabidopsis. *RNA* , 7(12), pp.1817–1832.

Brownlee, G.G., 1971. Sequence of 6S RNA of E. coli. *Nature*, 229(5), pp.147–149.

Burge, S.W. et al., 2013. Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, 41(Database issue), pp.D226–32.

Carthew, R.W. & Sontheimer, E.J., 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4), pp.642–655.

Cech, T.R. & Steitz, J.A., 2014. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*, 157(1), pp.77–94.

Chamary, J.V. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9), p.R75.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Chen, X., 2008. MicroRNA metabolism in plants. *Current topics in microbiology and immunology*, 320, pp.117–136.

Chou, C.-H. et al., 2016. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research*, 44(D1), pp.D239–47.

Clote, P. et al., 2005. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* , 11(5), pp.578–591.

Cohen, O. et al., 2016. Comparative transcriptomics across the prokaryotic tree of life. *Nucleic acids research*, 44(W1), pp.W46–W53.

Croucher, N.J. & Thomson, N.R., 2010. Studying bacterial transcriptomes using RNA-seq.

Current opinion in microbiology, 13(5), pp.619–624.

Cuperus, J.T., Fahlgren, N. & Carrington, J.C., 2011. Evolution and functional diversification of MIRNA genes. *The Plant cell*, 23(2), pp.431–442.

Darty, K., Denise, A. & Ponty, Y., 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* , 25(15), pp.1974–1975.

Darzacq, X. et al., 2002. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *The EMBO journal*, 21(11), pp.2746–2756.

Deutscher, M.P., 2006. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic acids research*, 34(2), pp.659–666.

Deutscher, M.P., 2003. Degradation of stable RNA in bacteria. *The Journal of biological chemistry*, 278(46), pp.45041–45044.

Dieterich, C. & Stadler, P.F., 2012. Computational biology of RNA interactions. *Wiley interdisciplinary reviews. RNA*, 4(1), pp.107–120.

Ding, Y. et al., 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485), pp.696–700.

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.

Do, C.B., Woods, D.A. & Batzoglou, S., 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* , 22(14), pp.e90–8.

Doshi, K.J. et al., 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5, p.105.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Eddy, S.R., 2006. Computational analysis of RNAs. *Cold Spring Harbor symposia on quantitative biology*, 71, pp.117–128.

Eddy, S.R. & Durbin, R., 1994. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11), pp.2079–2088.

Faner, M.A. & Feig, A.L., 2013. Identifying and characterizing Hfq-RNA interactions. *Methods* , 63(2), pp.144–159.

Freyhult, E.K., Bollback, J.P. & Gardner, P.P., 2007. Exploring genomic dark matter: a critical

assessment of the performance of homology search methods on noncoding RNA. *Genome research*, 17(1), pp.117–125.

Fürtig, B. et al., 2003. NMR spectroscopy of RNA. *Chembiochem: a European journal of chemical biology*, 4(10), pp.936–962.

Gardner, P.P. et al., 2009. Rfam: updates to the RNA families database. *Nucleic acids research*, 37(Database issue), pp.D136–40.

Gardner, P.P., Bateman, A. & Poole, A.M., 2010. SnoPatrol: how many snoRNA genes are there? *Journal of biology*, 9(1), p.4.

Gardner, P.P. & Eldai, H., 2015. Annotating RNA motifs in sequences and alignments. *Nucleic acids research*, 43(2), pp.691–698.

Gardner, P.P. & Giegerich, R., 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC bioinformatics*, 5(1), p.18.

Gardner, P.P., Wilm, A. & Washietl, S., 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic acids research*, 33(8), pp.2433–2439.

Gerlach, W. & Giegerich, R., 2006. GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics* , 22(6), pp.762–764.

Giannoukos, G. et al., 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome biology*, 13(3), p.R23.

Gilbert, W., 1986. Origin of life: The RNA world. *Nature*, 319(6055).

Gottesman, S., 2004. The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annual review of microbiology*, 58, pp.303–328.

Gou, L.-T. et al., 2015. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell research*, 25(2), p.266.

Gruber, A.R. et al., 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic acids research*, 35(Web Server issue), pp.W335–8.

Güell, M. et al., 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature reviews. Microbiology*, 9(9), pp.658–669.

Güell, M. et al., 2009. Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957), pp.1268–1271.

Gu, W., Zhou, T. & Wilke, C.O., 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS computational biology*, 6(2),

p.e1000664.

Hafner, M. et al., 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1), pp.129–141.

Hammer, B.K. & Bassler, B.L., 2007. Regulatory small RNAs circumvent the conventional quorum sensing pathway in pandemic Vibrio cholerae. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27), pp.11145–11149.

Harrow, J. et al., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9), pp.1760–1774.

Higgs, P.G. & Lehman, N., 2015. The RNA World: molecular cooperation at the origins of life. *Nature reviews. Genetics*, 16(1), pp.7–17.

Hindley, J., 1967. Fractionation of 32P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting. *Journal of molecular biology*, 30(1), pp.125–136.

Hodas, N.O. & Aalberts, D.P., 2004. Efficient computation of optimal oligo–RNA binding. *Nucleic acids research*, 32(22), pp.6636–6642.

Hoeppner, M.P., Gardner, P.P. & Poole, A.M., 2012. Comparative Analysis of RNA Families Reveals Distinct Repertoires for Each Domain of Life C. O. Wilke, ed. *PLoS computational biology*, 8(11), p.e1002752.

Hofacker, I.L. et al., 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte fuer Chemie*, 125(2), pp.167–188.

Hofacker, I.L., Fekete, M. & Stadler, P.F., 2002. Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, 319(5), pp.1059–1066.

Holmqvist, E. et al., 2016. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *The EMBO journal*, (35), pp.991–1011.

Holmqvist, E. & Vogel, J., 2013. A small RNA serving both the Hfq and CsrA regulons. *Genes & development*, 27(10), pp.1073–1078.

Hurst, L.D. & Merchant, A.R., 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings. Biological sciences / The Royal Society*, 268(1466), pp.493–497.

Hüttenhofer, A., Brosius, J. & Bachellerie, J.P., 2002. RNomics: identification and function of small, non-messenger RNAs. *Current opinion in chemical biology*, 6(6), pp.835–843.

Hüttenhofer, A. & Vogel, J., 2006. Experimental approaches to identify non-coding RNAs. *Nucleic acids research*, 34(2), pp.635–646.

Hyatt, D. et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, p.119.

Ingolia, N.T. et al., 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924), pp.218–223.

Jacob, F. & Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3, pp.318–356.

John, B. et al., 2004. Human MicroRNA targets. *PLoS biology*, 2(11), p.e363.

Jørgensen, M.G. et al., 2013. Dual function of the McaS small RNA in controlling biofilm formation. *Genes & development*, 27(10), pp.1132–1145.

Kadri, S., Hinman, V. & Benos, P.V., 2009. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC bioinformatics*, 10 Suppl 1, p.S35.

Kapranov, P. et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830), pp.1484–1488.

Karijolich, J. & Yu, Y.-T., 2010. Spliceosomal snRNA modifications and their function. *RNA biology*, 7(2), pp.192–204.

Katiyar, A. et al., 2012. Identification of miRNAs in sorghum by using bioinformatics approach. *Plant signaling & behavior*, 7(2), pp.246–259.

Kato, Y. et al., 2010. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* , 26(18), pp.i460–6.

Kehr, S. et al., 2011. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics* , 27(2), pp.279–280.

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. *Genome research*, 12(4), pp.656–664.

Kery, M.B. et al., 2014. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic acids research*, 42(Web Server issue), pp.W124–9.

Kiss, T., 2002. Small Nucleolar RNAs. *Cell*, 109(2), pp.145–148.

Klattenhoff, C. & Theurkauf, W., 2008. Biogenesis and germline functions of piRNAs. *Development* , 135(1), pp.3–9.

Klein, R.J., Misulovin, Z. & Eddy, S.R., 2002. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11), pp.7542–7547.

Knudsen, B. & Hein, J., 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13), pp.3423–3428.

Knudsen, B. & Hein, J., 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* , 15(6), pp.446–454.

Krogh, A. et al., 1994. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology*, 235(5), pp.1501–1531.

Kudla, G. et al., 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* , 324(5924), pp.255–258.

Kuhn, D.E. et al., 2008. Experimental validation of miRNA targets. *Methods* , 44(1), pp.47–54.

Kung, J.T.Y., Colognori, D. & Lee, J.T., 2013. Long noncoding RNAs: past, present, and future. *Genetics*, 193(3), pp.651–669.

Lai, D. & Meyer, I.M., 2015. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic acids research*, (44), p.13.

Layton, D.M. & Bundschuh, R., 2005. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic acids research*, 33(2), pp.519–524.

Lee, J.T., Davidow, L.S. & Warshawsky, D., 1999. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature genetics*, 21(4), pp.400–404.

Lindgreen, S. et al., 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS computational biology*, 10(10), p.e1003907.

Lin, M.F., Jungreis, I. & Kellis, M., 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* , 27(13), pp.i275–82.

Lloréns-Rico, V. et al., 2016. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science advances*, 2(3), p.e1501363.

Loh, E. et al., 2009. A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. *Cell*, 139(4), pp.770–779.

Lorenz, R. et al., 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology: AMB*, 6, p.26.

Loughrey, D. et al., 2014. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic acids research*, (44), p.13.

Lu, Z. et al., 2016. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, 165(5), pp.1267–1279.

Lu, Z. & Chang, H.Y., 2016. Decoding the RNA structurome. *Current opinion in structural biology*, 36, pp.142–148.

Mathews, D.H., 2006. Revolutions in RNA secondary structure prediction. *Journal of molecular biology*, 359(3), pp.526–532.

Mathews, D.H. & Turner, D.H., 2006. Prediction of RNA secondary structure by free energy minimization. *Current opinion in structural biology*, 16(3), pp.270–278.

Mattick, J.S., 2004. RNA regulation: a new genetics? *Nature reviews. Genetics*, 5(4), pp.316–323.

Mattick, J.S., 2009. The genetic signatures of noncoding RNAs. *PLoS genetics*, 5(4), p.e1000459.

Mattick, J.S. & Rinn, J.L., 2015. Discovery and annotation of long noncoding RNAs. *Nature structural & molecular biology*, (22), pp.5–7.

Mello, C.C. & Conte, D., Jr, 2004. Revealing the world of RNA interference. *Nature*, 431(7006), pp.338–342.

Millar, A.A. & Waterhouse, P.M., 2005. Plant and animal microRNAs: similarities and differences. *Functional & integrative genomics*, 5(3), pp.129–135.

Mistry, J. et al., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41(12), p.e121.

Modi, S.R. et al., 2011. Functional characterization of bacterial sRNAs using a network biology approach. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp.15522–15527.

Mückstein, U. et al., 2006. Thermodynamics of RNA–RNA binding. *Bioinformatics* , 22(10), pp.1177–1182.

Nawrocki, E.P. et al., 2015. Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, 43(Database issue), pp.D130–7.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R., 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* , 25(10), pp.1335–1337.

Nussinov, R. et al., 1978. Algorithms for Loop Matchings. *SIAM journal on applied mathematics*, 35(1), pp.68–82.

Oğul, H. et al., 2011. A probabilistic approach to microRNA-target binding. *Biochemical and biophysical research communications*, 413(1), pp.111–115.

Ogul, H. et al., 2013. TRAINER: a general-purpose trainable short biosequence classifer. *Protein*

and peptide letters, 20(10), pp.1108–1114.

Omer, A.D. et al., 2000. Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465), pp.517–522.

Onoa, B. & Tinoco, I., Jr, 2004. RNA folding and unfolding. *Current opinion in structural biology*, 14(3), pp.374–379.

van Opijnen, T., Bodi, K.L. & Camilli, A., 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature methods*, 6(10), pp.767–772.

Pain, A. et al., 2015. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5), pp.509–513.

Park, C. et al., 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8), pp.E678–86.

Pauli, A., Valen, E. & Schier, A.F., 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 37(1), pp.103–112.

Pearson, W.R. & Lipman, D.J., 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8), pp.2444–2448.

Pedersen, J.S. et al., 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS computational biology*, 2(4), p.e33.

Pennisi, E., 2012. Genomics. ENCODE project writes eulogy for junk DNA. *Science*, 337(6099), pp.1159, 1161.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1), pp.32–42.

Prasse, D. et al., 2013. Regulatory RNAs in archaea: first target identification in Methanoarchaea. *Biochemical Society transactions*, 41(1), pp.344–349.

Rath, D. et al., 2015. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*, 117, pp.119–128.

Rehmsmeier, M. et al., 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* , 10(10), pp.1507–1517.

Reuter, J.S. & Mathews, D.H., 2010. RNAstructure: software for RNA secondary structure

prediction and analysis. *BMC bioinformatics*, 11, p.129.

Riley, K.J. & Steitz, J.A., 2013. The "Observer Effect" in Genome-wide Surveys of Protein-RNA Interactions. *Molecular cell*, 49(4), pp.601–604.

Rivas, E. & Eddy, S.R., 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics*, 2, p.8.

Rivas, E. & Eddy, S.R., 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* , 16(7), pp.583–605.

Schattner, P., 2002. Searching for RNA genes using base-composition statistics. *Nucleic acids research*, 30(9), pp.2076–2082.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , 30(14), pp.2068–2069.

Shao, Y. & Bassler, B.L., 2012. Quorum-sensing non-coding small RNAs use unique pairing regions to differentially control mRNA targets. *Molecular microbiology*, 83(3), pp.599–611.

Sharma, C.M. et al., 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286), pp.250–255.

Sharma, C.M. & Vogel, J., 2014. Differential RNA-seq: the approach behind and the biological insight gained. *Current opinion in microbiology*, 19, pp.97–105.

Sharma, E. et al., 2016. Global Mapping of Human RNA-RNA Interactions. *Molecular cell*, 62(4), pp.618–626.

Sharp, P.A., 2009. The centrality of RNA. *Cell*, 136(4), pp.577–580.

Staple, D.W. & Butcher, S.E., 2005. Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6), p.e213.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Tafer, H. et al., 2010. RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics* , 26(5), pp.610–616.

Talkish, J. et al., 2014. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* , 20(5), pp.713–720.

Tat, T.T. et al., 2016. Cotranslational microRNA mediated messenger RNA destabilization. *eLife*, 5(e12880). Available at: http://dx.doi.org/10.7554/eLife.12880.

Tinoco, I., Jr & Bustamante, C., 1999. How RNA folds. *Journal of molecular biology*, 293(2),

pp.271–281.

Turner, D.H., Sugimoto, N. & Freier, S.M., 1988. RNA structure prediction. *Annual review of biophysics and biophysical chemistry*, 17, pp.167–192.

Umu, S.U. et al., 2015. *Natural avoidance of stochastic mRNA:ncRNA interactions can be harnessed to control protein expression levels*, Available at: http://biorxiv.org/lookup/doi/10.1101/033613.

Underwood, J.G. et al., 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods*, 7(12), pp.995–1001.

Vogel, J., 2009. A rough guide to the non-coding RNA world of Salmonella. *Molecular microbiology*, 71(1), pp.1–11.

Vogel, J. & Luisi, B.F., 2011. Hfq and its constellation of RNA. *Nature reviews. Microbiology*, 9(8), pp.578–589.

Vogel, J. & Sharma, C.M., 2005. How to find small non-coding RNAs in bacteria. *Biological chemistry*, 386(12), pp.1219–1238.

Vogel, J. & Wagner, E.G.H., 2007. Target identification of small noncoding RNAs in bacteria. *Current opinion in microbiology*, 10(3), pp.262–270.

Wade, J.T. & Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature reviews. Microbiology*, 12(9), pp.647–653.

Wang, J. et al., 2016. sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic acids research*, 44(D1), pp.D248–53.

Wang, L. et al., 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6), p.e74.

Washietl, S. et al., 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* , 17(4), pp.578–594.

Washietl, S. & Hofacker, I.L., 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *Journal of molecular biology*, 342(1), pp.19–30.

Wassarman, K.M. & Storz, G., 2000. 6S RNA regulates E. coli RNA polymerase activity. *Cell*, 101(6), pp.613–623.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

Weinberg, Z. & Breaker, R.R., 2011. R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC bioinformatics*, 12, p.3.

Wenzel, A., Akbaşli, E. & Gorodkin, J., 2012. RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* , 28(21), pp.2738–2746.

Wheeler, T.J. & Eddy, S.R., 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* , 29(19), pp.2487–2489.

Wilbert, M.L. & Yeo, G.W., 2011. Genome-wide approaches in the study of microRNA biology. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(5), pp.491–512.

Workman, C. & Krogh, A., 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24), pp.4816–4822.

Wright, P.R. et al., 2013. Comparative genomics boosts target prediction for bacterial small RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 110(37), pp.E3487–96.

Wuchty, S. et al., 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2), pp.145–165.

Wu, Y. et al., 2011. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*, 12, p.107.

Xia, T. et al., 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42), pp.14719–14735.

Xue, C. et al., 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, 6, p.310.

Yan, B., Wang, Z.-H. & Guo, J.-T., 2012. The research strategies for probing the function of long noncoding RNAs. *Genomics*, 99(2), pp.76–80.

Yang, Q. et al., 2013. MicroRNA and piRNA profiles in normal human testis detected by next generation sequencing. *PloS one*, 8(6), p.e66809.

Yang, Y., Wang, Y.-P. & Li, K.-B., 2008. MiRTif: a support vector machine-based microRNA target interaction filter. *BMC bioinformatics*, 9 Suppl 12, p.S4.

Zhang, J. & Ferré-D'Amaré, A.R., 2015. Structure and mechanism of the T-box riboswitches. *Wiley interdisciplinary reviews. RNA*, 6(4), pp.419–433.

Zhang, M.Q., 2002. Computational prediction of eukaryotic protein-coding genes. *Nature reviews. Genetics*, 3(9), pp.698–709.

Zhang, P. et al., 2015. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse

testes. *Cell research*, 25(2), pp.193–207.

Zhao, Y. et al., 2016. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic acids research*, 44(D1), pp.D203–8.

Zuker, M., Mathews, D.H. & Turner, D.H., 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In *RNA Biochemistry and Biotechnology*. NATO Science Series. Springer Netherlands, pp. 11–43.

Zuker, M. & Sankoff, D., 1984. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4), pp.591–621.

Zuker, M. & Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1), pp.133–148.

Zur, H. & Tuller, T., 2012. Strong association between mRNA folding strength and protein abundance in S. cerevisiae. *EMBO reports*, 13(3), pp.272–277.

CHAPTER II - An RNA Encyclopedia for Bacteria and Archaea

Bacteria and archaea are already known to contain a rich repertoire of RNA elements such as riboswitches (Waters & Storz 2009; Breaker 2012), transcription terminators (Gardner et al. 2011; Santangelo & Artsimovitch 2011), CRISPR RNAs (Barrangou et al. 2007; Bhaya et al. 2011; Rath et al. 2015), sRNAs (Waters & Storz 2009; Babski et al. 2014), thermoregulators (Narberhaus & Waldminghaus 2006) and snoRNAs of archaea (Omer et al. 2000; Gardner et al. 2010).

Advances in sequencing technology have also revealed that non-canonical transcripts (i.e. usually antisense, non-coding) and transcripts without any protein product are common in prokaryotes (i.e. genomic dark matter) (Wade & Grainger 2014; Lindgreen et al. 2014; Croucher & Thomson 2010; Güell et al. 2009; Chen et al. 2016). On the other hand, there has been a lot of debate in the RNA community (for both eukaryotes and prokaryotes) about whether expression equals function, and what level of evidence is required to illuminate the 'genomic dark matter' (Palazzo & Gregory 2014; Wade & Grainger 2014; Pauli et al. 2015; Clark et al. 2011; van Bakel et al. 2010; Lloréns-Rico et al. 2016; Kapranov et al. 2007), and the products of transcription can be results of a simple biological or experimental noise (Wade & Grainger 2014; van Bakel et al. 2010; Goldman et al. 2009; Baboo & Cook 2014; Lloréns-Rico et al. 2016; Cohen et al. 2016).

The aim of the following study is to examine prokaryotic transcriptomes for novel ncRNA genes using available high-throughput sequencing data. To achieve this goal, we collected the publicly available transcriptome data of prokaryotes from various studies (as of August 2013). In total we processed data spanning 37 strains of Bacteria-Archaea and 413 RNA-seq experiments. We found that (1) ncRNAs are highly abundant in prokaryotic genomes, (2) current RNA-seq data is heavily biased towards some strains (e.g. model organisms, pathogens etc.) and (3) most of the

ncRNAs are evolutionarily young (Lindgreen et al. 2014).

Therefore, we have proposed a phylogeny-informed approach to detect (functional) novel ncRNA genes; otherwise, it is harder to characterize unknown ncRNA transcripts and separate them from transcriptional noise. In other words, RNA-seq experiments must sample the strains inside an optimal phylogenetic zone, which we colloquially refer to as 'the Goldilocks zone', to invoke the essential comparative power for an effective ncRNA detection.

I contributed to the project reported in the following pages as a joint-first author and was part of each stage. I made a major contribution to the bioinformatic methodology, data analysis and acquisition. I also contributed to manuscript writing, particularly with figure creation, and summary statistical analysis. I am continuing to work on the next stage of the project 'An RNA Encyclopedia of Bacteria and Archaea II (AREBA-II)' (other than my advisors) and I am maintaining the associated data repository. In AREBA-II, our plan is to apply the Goldilocks zone sampling to selected taxonomic groups (which I will summarize in Chapter V).

REFERENCES

Baboo, S. & Cook, P.R., 2014. "Dark matter" worlds of unstable RNA and protein. *Nucleus* , 5(4), pp.281–286.

Babski, J. et al., 2014. Small regulatory RNAs in Archaea. *RNA biology*, 11(5), pp.484–493.

van Bakel, H. et al., 2010. Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLoS biology*, 8(5), p.e1000371.

Barrangou, R. et al., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), pp.1709–1712.

Bhaya, D., Davison, M. & Barrangou, R., 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, pp.273–297.

Breaker, R.R., 2012. Riboswitches and the RNA world. *Cold Spring Harbor perspectives in biology*, 4(2), p.a003566.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium:

Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Clark, M.B. et al., 2011. The reality of pervasive transcription. *PLoS biology*, 9(7), p.e1000625; discussion e1001102.

Cohen, O. et al., 2016. Comparative transcriptomics across the prokaryotic tree of life. *Nucleic acids research*, 44(W1), p.Pp. W46–W53.

Croucher, N.J. & Thomson, N.R., 2010. Studying bacterial transcriptomes using RNA-seq. *Current opinion in microbiology*, 13(5), pp.619–624.

Gardner, P.P. et al., 2011. RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic acids research*, 39(14), pp.5845–5852.

Gardner, P.P., Bateman, A. & Poole, A.M., 2010. SnoPatrol: how many snoRNA genes are there? *Journal of biology*, 9(1), p.4.

Goldman, S.R., Ebright, R.H. & Nickels, B.E., 2009. Direct detection of abortive RNA transcripts in vivo. *Science*, 324(5929), pp.927–928.

Güell, M. et al., 2009. Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957), pp.1268–1271.

Kapranov, P. et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830), pp.1484–1488.

Lindgreen, S. et al., 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS computational biology*, 10(10), p.e1003907.

Lloréns-Rico, V. et al., 2016. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science advances*, 2(3), p.e1501363.

Narberhaus, F. & Waldminghaus, T., 2006. RNA thermometers. *FEMS microbiology*, 30(1), pp.3–16.

Omer, A.D. et al., 2000. Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465), pp.517–522.

Palazzo, A.F. & Gregory, T.R., 2014. The case for junk DNA. *PLoS genetics*, 10(5), p.e1004351.

Pauli, A., Valen, E. & Schier, A.F., 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 37(1), pp.103–112.

Rath, D. et al., 2015. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*, 117, pp.119–128.

Santangelo, T.J. & Artsimovitch, I., 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nature reviews. Microbiology*, 9(5), pp.319–329.

Wade, J.T. & Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature reviews. Microbiology*, 12(9), pp.647–653.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Robust Identification of Noncoding RNA from Transcriptomes Requires Phylogenetically-Informed Sampling

Stinus Lindgreen[1,2☯], Sinan Uğur Umu[2,3☯], Alicia Sook-Wei Lai[2], Hisham Eldai[2], Wenting Liu[2], Stephanie McGimpsey[2], Nicole E. Wheeler[2], Patrick J. Biggs[4,5], Nick R. Thomson[6], Lars Barquist[6,7], Anthony M. Poole[2,3,5*], Paul P. Gardner[2,3*]

1 Department of Biology, University of Copenhagen, Copenhagen, Denmark, **2** School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, **3** Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand, **4** Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Palmerston North, New Zealand, **5** Allan Wilson Centre for Molecular Ecology & Evolution, Massey University, Palmerston North, New Zealand, **6** Pathogen Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, **7** Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany

Abstract

Noncoding RNAs are integral to a wide range of biological processes, including translation, gene regulation, host-pathogen interactions and environmental sensing. While genomics is now a mature field, our capacity to identify noncoding RNA elements in bacterial and archaeal genomes is hampered by the difficulty of *de novo* identification. The emergence of new technologies for characterizing transcriptome outputs, notably RNA-seq, are improving noncoding RNA identification and expression quantification. However, a major challenge is to robustly distinguish functional outputs from transcriptional noise. To establish whether annotation of existing transcriptome data has effectively captured all functional outputs, we analysed over 400 publicly available RNA-seq datasets spanning 37 different Archaea and Bacteria. Using comparative tools, we identify close to a thousand highly-expressed candidate noncoding RNAs. However, our analyses reveal that capacity to identify noncoding RNA outputs is strongly dependent on phylogenetic sampling. Surprisingly, and in stark contrast to protein-coding genes, the phylogenetic window for effective use of comparative methods is perversely narrow: aggregating public datasets only produced one phylogenetic cluster where these tools could be used to robustly separate unannotated noncoding RNAs from a null hypothesis of transcriptional noise. Our results show that for the full potential of transcriptomics data to be realized, a change in experimental design is paramount: effective transcriptomics requires phylogeny-aware sampling.

Citation: Lindgreen S, Umu SU, Lai AS-W, Eldai H, Liu W, et al. (2014) Robust Identification of Noncoding RNA from Transcriptomes Requires Phylogenetically-Informed Sampling. PLoS Comput Biol 10(10): e1003907. doi:10.1371/journal.pcbi.1003907

Editor: Kevin Chen, Rutgers University, United States of America

Received July 22, 2014; **Accepted** September 11, 2014; **Published** October 30, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files and on GitHub: https://github.com/UCanCompBio/AREBA

Funding: SL is supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme. SUU is supported by a Biomolecular Interaction Centre and Bluefern Supercomputing Facility joint PhD Scholarship from the University of Canterbury. LB is supported by a Research Fellowship from the Alexander von Humboldt Stiftung/Foundation. NRT is supported by the Wellcome Trust (grant number 098051). AMP & PPG are both supported by Rutherford Discovery Fellowships, administered by the Royal Society of New Zealand. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: anthony.poole@canterbury.ac.nz (AMP); paul.gardner@canterbury.ac.nz (PPG)

☯ These authors contributed equally to this work.

Introduction

Genome sequencing has transformed microbiology, offering unprecedented insight into the physiology, biochemistry, and genetics of Bacteria and Archaea [1–4]. Equally, careful examination of transcriptional outputs has revealed that bacterial and archaeal transcriptomes are remarkably complex [5]. Roles for RNA include regulation, post-transcriptional modification and genome defense processes [6–10]. However, our view of the microbial RNA world still derives from a narrow sampling of microbial diversity [11]. Additional bias comes from the fact that many microbes are not readily culturable [12]. The development of metagenomics and initiatives such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project have sought to redress these biases, generating genomes spanning undersampled regions of the bacterial and archaeal phylogeny [1], and sequencing uncultured or unculturable species through metagenomics [2,13–16].

A further source of bias in our genome-informed view of microbes derives from a protein-centric approach to genome annotation. The majority of genome sequences deposited in public databases carry limited annotation of noncoding RNAs and cis-regulatory elements, yet it is rapidly becoming clear that RNA is essential to our understanding of molecular functioning in microbes [17].

The paucity of annotations is understandable, as RNA gene annotation is non-trivial [18,19]. However, the increasing number of roles for RNAs uncovered through experimental and bioinformatic studies make illuminating this "dark matter"

Author Summary

We have analysed more than 400 public transcriptomes, generated using RNA-seq, from almost 40 strains of Bacteria and Archaea. We discovered that the capacity to identify noncoding RNA outputs from this data is strongly dependent on phylogenetic sampling. Our results show that, for the full potential of transcriptomics data as a discovery tool to be realized, a change in experimental design is critical: effective comparative transcriptomics requires phylogeny-aware sampling. We also examined how comparative transcriptomics experiments can be used to effectively identify RNA elements. We find that, for RNA element discovery, a phylogeny-informed sampling approach is more effective than analyses of individual species. Phylogeny-informed sampling reveals a narrow 'Goldilocks Zone' (where species are not too similar and not too divergent) for RNA identification using clusters of related species. In stark contrast to protein-coding genes, not only is the phylogenetic window for the effective use of comparative methods for noncoding RNA identification perversely narrow, but few existing datasets sit within this Goldilocks Zone: by aggregating public datasets, we were only able to create one phylogenetic cluster where comparative tools could be used to confidently separate unannotated noncoding RNAs from transcriptional noise.

all the more urgent. Among the remarkable discoveries made are: riboswitch-mediated regulation [9,20], transcriptional termination by RNA elements [21–23], identification of novel natural catalytic RNAs [24–27], CRISPR-mediated acquired immunity [28,29], temperature-dependent gene regulation [30,31], and sno-like RNAs in Archaea [32–34]. The Rfam database [22,35] provides a valuable platform for collating and characterising these and other families of noncoding RNA. However, a recent comparative analysis [36] revealed that fewer than 7% of RNA families within Bacteria and less than 19% in Archaea show a broad phylogenetic distribution (that is, presence in at least 50% of sequenced phyla). Crucially, that analysis revealed that underlying genome sequencing biases were a major contributor to this pattern, and that the wider genomic sampling provided by the GEBA dataset [1] did help improve identification of broadly-conserved RNA families [36]. Tools such as RNA-seq [37] and transposon insertion sequencing [38–40] promise to complement comparative genomics tools for RNA family discovery, and it may be possible to use a mix of data types in the identification of RNA elements. However, to date, no systematic analysis of available data has been undertaken, suggesting ncRNAs may be hidden in the deluge of published data.

We have therefore assessed the value of RNA-seq data for identification of unannotated non-coding and cis-regulatory RNA elements in bacterial and archaeal genomes. We show that numerous, hitherto uncharacterised, expressed RNA families are lurking in publicly available RNA-seq datasets. We find that poor sequence conservation for RNA families limits the capacity to identify evolutionarily conserved, expressed ncRNAs from existing genomic and transcriptomic data. Our results suggest that maximising phylogenetic distance, a sampling strategy effective for identification of novel protein families [1,2], is not the most effective strategy for ncRNA identification. Instead, our results show that, for RNA element identification, sequencing clusters of related microbes will generate the greatest benefit.

Results

Non-coding RNA elements dominate bacterial and archaeal transcriptional profiles

To assess the relative contribution of noncoding RNAs and protein-coding genes to transcriptional output, we collected all publicly-available bacterial and archaeal RNA-seq datasets (available as of August 2013), spanning 37 species/strains and 413 datasets. For all datasets, we supplemented publicly available genome annotations with screening for additional loci against the Pfam and Rfam databases [22,35,41,42], followed by manual identification of expressed unannotated regions that have previously been dubbed RNAs of Unknown Function (RUFs) [43]. This latter annotation yielded 922 expressed RUFs.

We next examined the relative abundance of transcripts within each RNA-seq dataset, yielding an expression rank for individual transcripts. This analysis reveals that most transcriptomes are dominated by highly expressed non-coding RNA outputs (Figure 1) (P-value $<<0.0001$, Chi-square test of observed vs. expected ratios and Fisher's Exact test on the counts). In addition to well-characterised RNAs (rRNA, tRNA, tmRNA, RNase P RNA, SRP RNA, 6S and sno-like sRNAs), and known cis-regulatory elements (riboswitches, leaders and thermosensors - Table S1), the top 50 most abundant transcriptional outputs (Figure 1) across the 32 Bacteria and 5 Archaea in our dataset included a total of 308 RUFs.

Comparative analyses reveal that highly expressed transcripts are often poorly conserved

To assess whether highly expressed RUFs possess features commonly associated with function, we employed three criteria: 1) evolutionary conservation, 2) conservation of secondary structure, 3) evidence of expression in more than one RNA-seq dataset. For this analysis, we compared and ranked transcriptional outputs across species/strains (see Methods for details). Based on the relative rank across RNA-seq datasets and the maximum phylogenetic distance observed across all genomes, each transcript was classified as high, medium or low expression, and high, medium or low conservation. This yielded a set of highly expressed transcripts consisting of 162 Rfam families, 568 RUFs and 1429 Pfam families. As expected [44–46], conserved, highly expressed outputs are dominated by protein-coding transcripts (Figure 2 B&C). In contrast, transcripts that are highly expressed but poorly conserved are primarily RUFs (Figure 2A). Of the 568 RUFs identified, only 25 are supported by all three conservative criteria (conservation, secondary structure and expression) (Figure 2D), a further 138 RUFs are supported by two criteria (Figure 2D). Consequently, on these criteria, the vast majority of RUFs appear indistinguishable from transcriptional noise. However, as these RUFs are among the most highly expressed transcripts in public RNA-seq data, we next considered whether our criteria were sufficiently discriminatory to identify functional RNAs. It is well established that not all functional RNAs exhibit conserved secondary structure – antisense base pairing with a target is common, and does not require intramolecular folding [47]. This indicates that criterion 2 will apply to some, but not all functional RNA elements. Criteria 1 and 3 both derive from comparative analysis: criterion 1 requires an expressed RUF to be conserved in some other genome, while criterion 2 requires an expressed RUF to be expressed in another of the datasets in our study. We therefore sought to examine how effective our comparative analyses are given that the available data represent a small sample (transcriptomes from 37 strains) and given that biases in genome

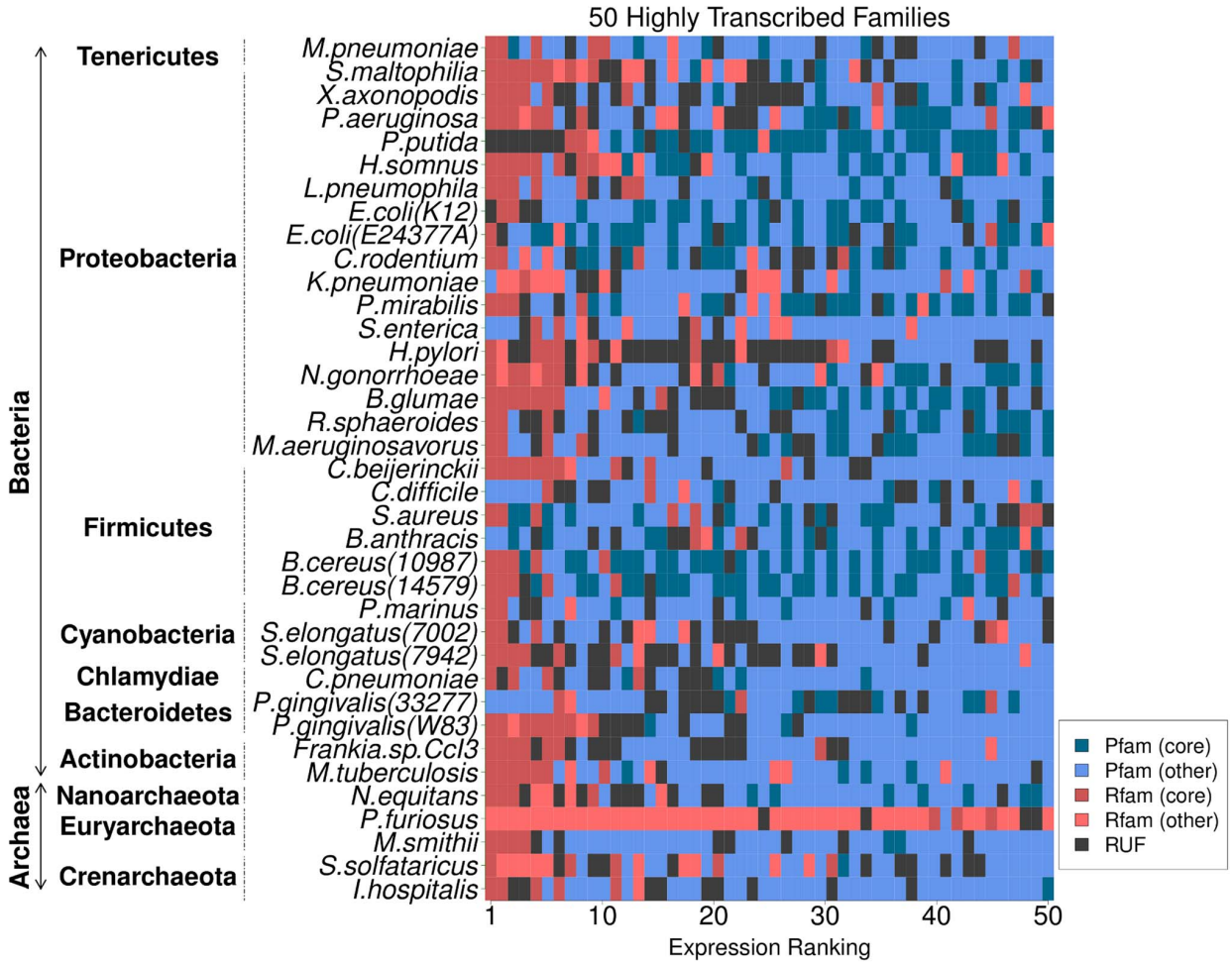

Figure 1. Identification of transcribed elements across publicly-available RNA-seq data. Non-coding RNA elements show high expression across transcriptomes. Both annotated Rfam families (red - core Rfam families (see Methods) are dark red, all others are light red) and expressed RUFs (black) are among the highest expressed outputs in transcriptomes (blue - core Pfam families (see Methods) are dark blue, all others are light blue). For each strain we generated relative rankings of expression spanning protein coding genes, RNA genes and candidate RUFs. Accurately estimating expression levels from read depths is confounded by a number of factors (e.g. sample preparation, overall sequencing depths, rRNA depletion, etc.). For consistency, we have ranked genes for each strain and compared rankings instead of comparing the read depths directly between strains. For a given strain, the annotated genes were ranked based on the median read depth of the annotated region. RUFs were manually picked by masking out annotated genes and selecting regions showing evidence of expression by inspecting read depth across the genome. This yielded 844 gene candidate sequences in Bacteria and 78 in Archaea. The plot contains the 50 most highly expressed elements for each strain/species.
doi:10.1371/journal.pcbi.1003907.g001

sampling across bacterial and archaeal diversity impact comparative analysis of RNAs [36].

Comparative analysis reveals a 'Goldilocks Zone' for ncRNA identification

Effective comparative analysis requires appropriate phylogenetic distances between species under investigation [48]. For discovery of protein-coding gene families, maximising phylogenetic diversity across the tree of life has proven very effective [1,2]. For non-coding RNA, underlying biases in genome sampling do affect the assessment of ncRNA conservation, and adding phylogenetic diversity improves the identification of broadly conserved ncRNA families [36]. However, few ncRNAs appear conserved across broad evolutionary distances [36]. We have therefore considered how species selection impacts comparative analysis as a tool for the identification of conserved ncRNAs.

To assess the effect of strain selection on our capacity to identify RNA families using comparative analysis, we first generated F84 phylogenetic distances between 2562 bacterial strains and 154 archaeal strains using SSU rRNA sequences from each strain (see Methods for details). Next, for each Rfam RNA family and Pfam protein family, we identified the maximum phylogenetic distance between any two species/strains that encode a given family. We then calculated the fraction of conserved RNA and protein families for a given phylogenetic distance.

This reveals a dramatic difference in evolutionary conservation of Rfam and Pfam families (Figure 3). While 80% of protein families are still conserved at the broad evolutionary distances that separate Bacteria and Archaea, the phylogenetic distance at which 80% of RNA families are conserved lies somewhere between the taxonomic levels of genus and family (Figure 3). The explanation for this rapid decay of RNA family conservation across long

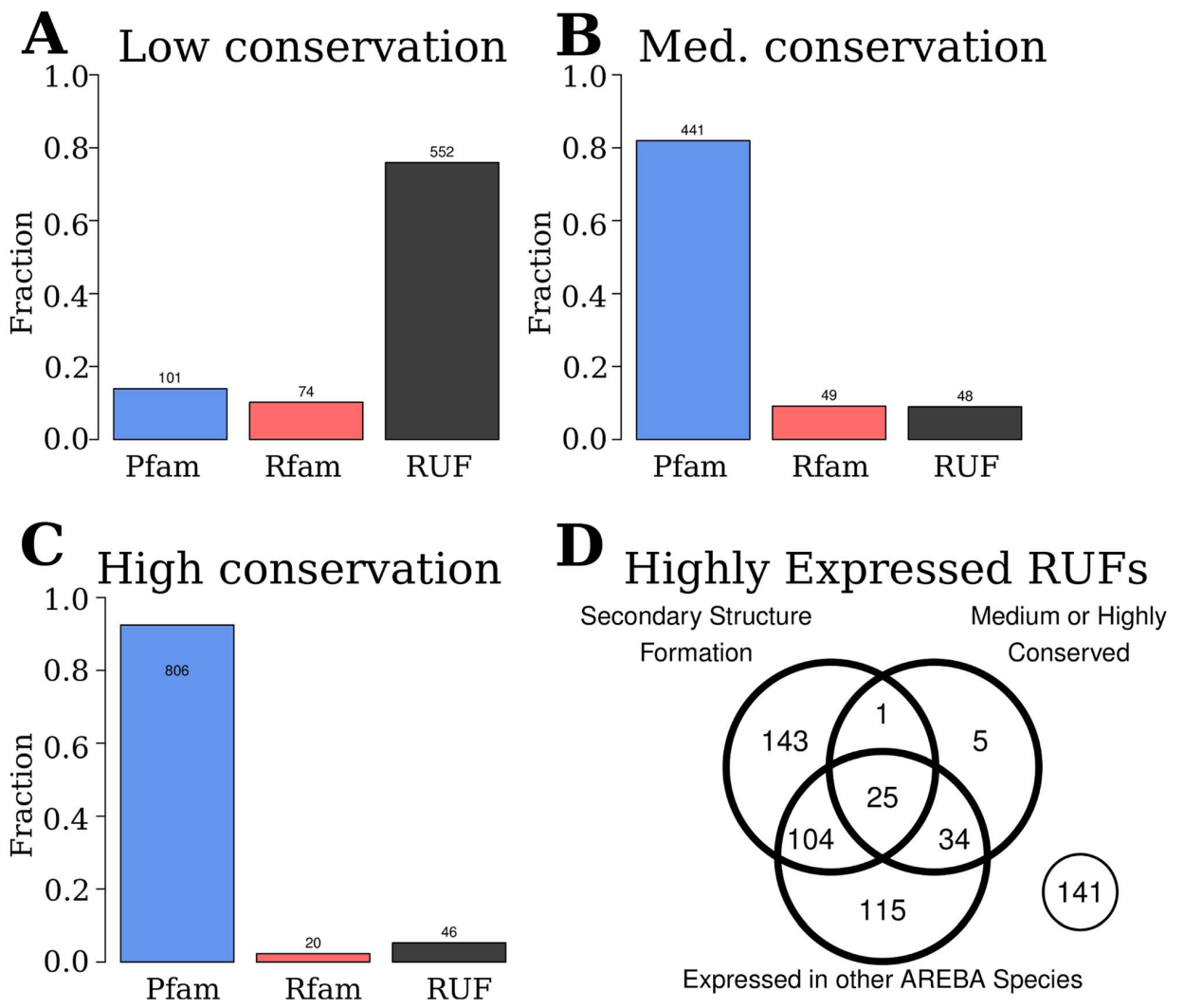

Figure 2. Many ncRNAs and RUFs are highly expressed but show limited conservation across represented strains/species. A–C: We have defined the "family conservation" for Pfam, Rfam and RUFs based upon the maximum phylogenetic distance (using structural SSU rRNA alignments) between any two strains hosting the family. We have divided the highly expressed transcripts (ranks 1–204) into Low, Medium and High conservation groups based on the lower-quartile, inter-quartile range and the upper-quartile of the family conservation measure (see Methods for further details). Both the known Rfam families and the RUFs identified in this analysis are often highly expressed transcripts. In contrast to protein-coding transcripts (blue), where highly-expressed transcripts are well-conserved, the opposite is true of many non-coding RNA elements (Rfam, red; RUFs, black). Notably, the greatest proportion of highly expressed Rfam-annotated RNA elements show a narrow evolutionary distribution. This is also reflected in the RUFs identified in this study. **D:** Venn diagram of the 568 highly expressed RUFs. Each RUF was analysed to look for evidence of secondary structure formation, level of conservation, and evidence of expression in at least one other RNA-seq dataset. All RUFs showing expression in other strains/species are conserved in at least two strains/species, so the figure also shows that 219 highly expressed RUFs are conserved across a limited phylogenetic distance only.
doi:10.1371/journal.pcbi.1003907.g002

evolutionary time-scales is likely to be a combination of the limited abilities of existing bioinformatic tools to correctly align RNA sequences [49] and rapid turnover of non-coding RNAs during evolution [36].

These results in turn indicate that appropriate evolutionary distances for optimal comparative analysis differ greatly for protein- and RNA-coding genes. Figure 3 confirms the utility of the GEBA sampling strategy [1,2] for protein-coding gene identification, since maximising phylogenetic diversity permits effective identification of conserved protein-coding genes. In contrast, at the largest phylogenetic distances, less than 40% of the RNA families are amenable to comparative analysis. These results define a 'Goldilocks Zone' (an evolutionary distance neither too close nor too distant) for ncRNA analysis through comparative analysis.

In order to assess the potential for existing RNA-seq data to be used for ncRNA analysis, we mapped the pairwise distances between strains covered by the RNA-seq datasets in this study. Of the 506 possible pairs (excluding Bacteria vs Archaea), only 11 are in the Goldilocks Zone for RNA (phylogenetic distance between 0.0118 and 0.0542) covering 9 species/strains. While five pairs of datasets are 'too hot' (i.e. too close phylogenetically), the remaining 490 comparisons are 'too cold' for effective comparative RNA analysis (Figure 3). The datasets in the Goldilocks Zone span three

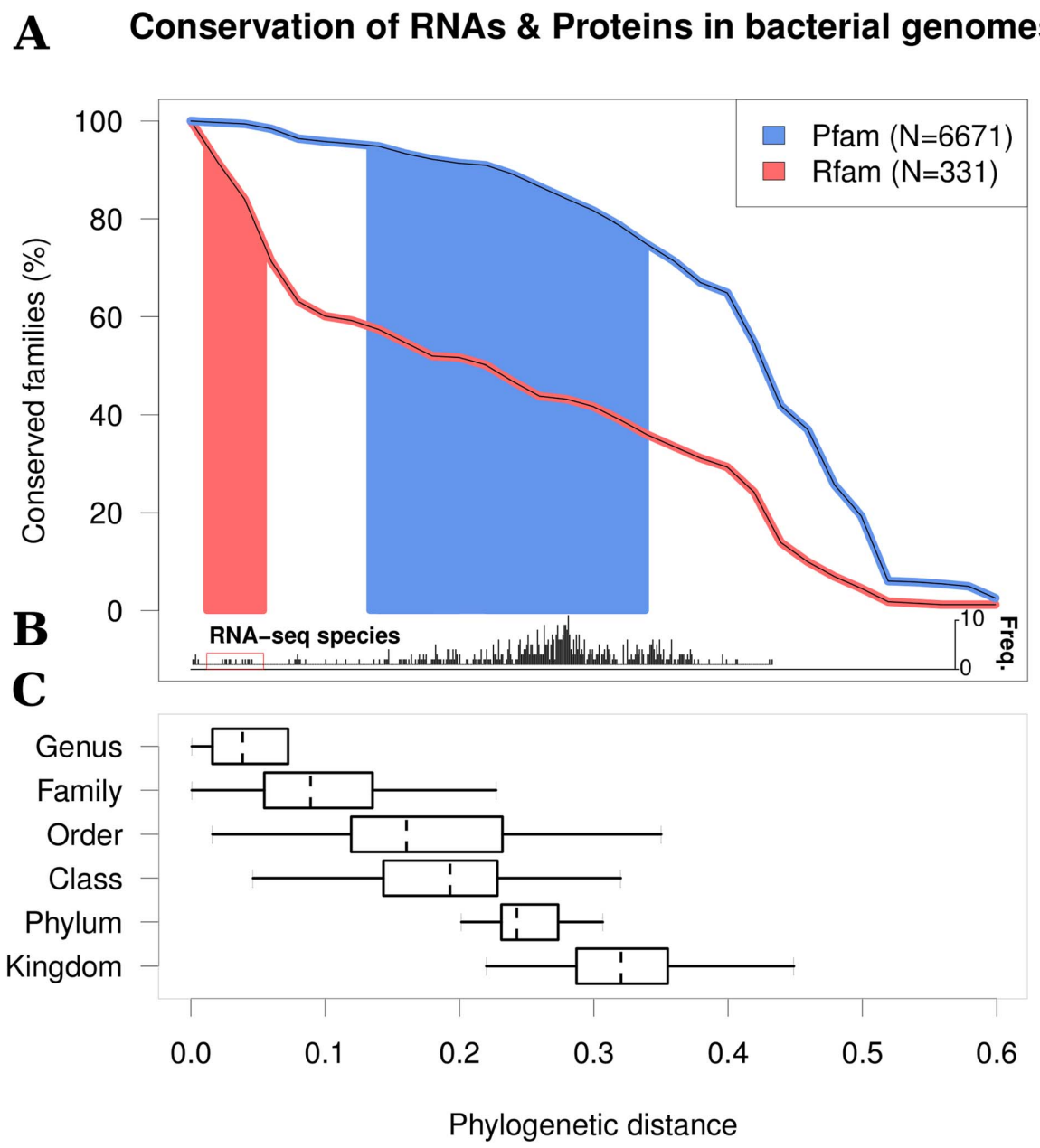

Figure 3. Conservation of protein and RNA families. All of the available full length Bacterial and Archaeal genomes were annotated using Rfam and Pfam models. For each Pfam/Rfam family, RNA-seq species or taxonomic group the "phylogenetic distance" is calculated using the maximum SSU rRNA F84 distance (see Methods for details). **A.** For the Pfam and the Rfam families we compare the levels of conservation as a function of phylogenetic distance using annotations of 2,562 bacterial genomes. E.g. $\approx 60\%$ of RNA families are conserved between species from the same family, whereas $>90\%$ of protein families are conserved within the same taxonomic range. **B.** The barplot shows the distribution of all pairwise distances between the RNA-seq datasets. Eleven pairs (boxed) are in the Goldilocks Zone (See Figure 4 for further analysis). **C.** The ranges of phylogenetic distances for comparing species from different taxonomic groups.
doi:10.1371/journal.pcbi.1003907.g003

distinct clades covering five Enterobacteria, three Pseudomanada, and two Xanthomonada (Figure 4).

We next calculated the percentage of conserved RUFs for all Enterobacterial strain pairs. On average, 83% of RUFs are conserved across the Goldilocks Zone. The two *E. coli* strains are extremely similar, and share 99% of their RUFs, suggesting that these strains are too similar for us to robustly separate expression of *bona fide* RNAs from noise. While these outputs could be genuine RNAs, these strains are in the 'too hot' region, meaning if everything is conserved, comparative power is lost. In contrast, only 12% of RUFs are conserved between strains/species pairs in the 'too cold' region (spanning clades; Figure 4) and of the 197 RUFs found through comparative analysis of transcriptomes within the Goldilocks Zone, only 19 show evidence of expression

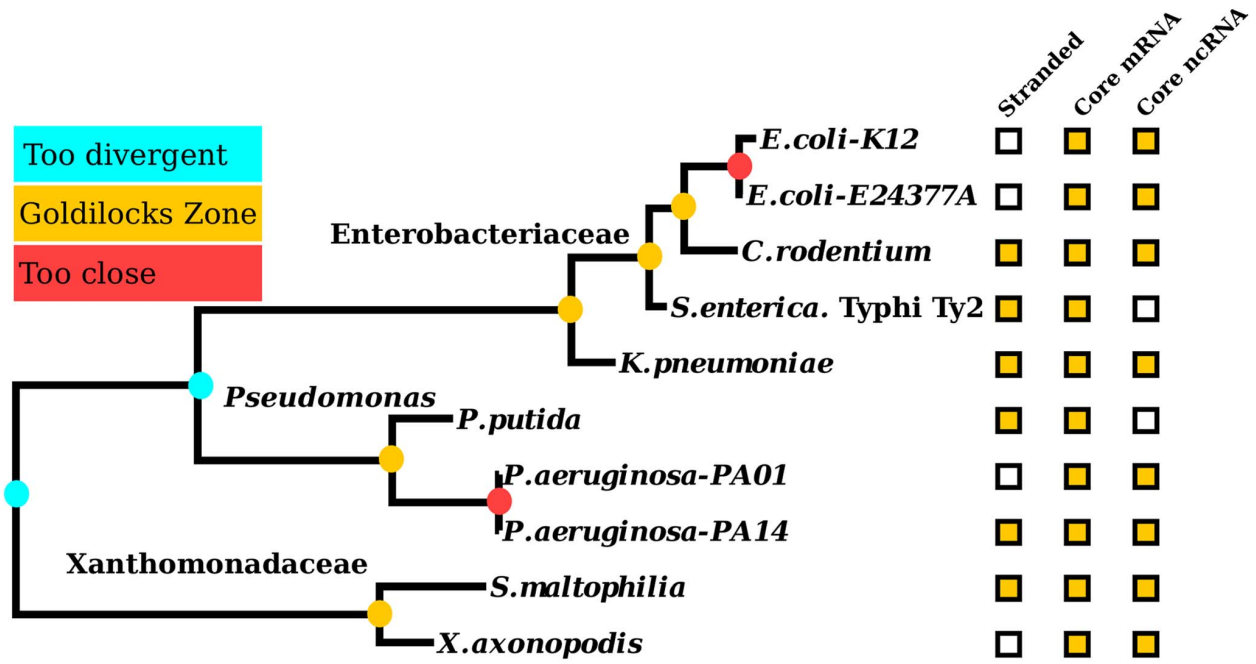

Figure 4. Public RNA-seq datasets that lie in the Goldilocks Zone. Ten strains with corresponding, publicly available RNA-seq data and phylogenetic distances in the Goldilocks Zone (Figure 3) have been identified. The maximum likelihood tree from a SSU rRNA alignment shows the relationships between these taxa. They fall into three clades, containing members of the families: Enterobacteriaceae and Xanthomonadaceae, and the genus: *Pseudomonas*. The nodes connecting taxa within the Goldilocks Zone are coloured gold, taxa that are too close are coloured red and those that are too divergent are coloured cyan. Each strain is annotated with gold boxes where there was stranded information, or if the majority of core mRNAs and ncRNAs (see Methods) were expressed (see Table S3 for the raw data).
doi:10.1371/journal.pcbi.1003907.g004

in another transcriptome outside of this zone. This suggests that the low number of RUFs from Figure 2D showing both conservation and expression is primarily a consequence of limited sampling. That said, mining RNA-seq data within the Goldilocks Zone permits a higher confidence in the identification of novel ncRNAs. Three examples of this are illustrated in Figure 5. These RUFs exhibit sequence and secondary structure conservation and are expressed at high levels across multiple Goldilocks Zone transcriptomes.

In summary, the Goldilocks Zone for RNA is surprisingly narrow, and suggests that optimal strain selection for RNA comparative analyses should comprise strains of the same species, members of the same genus, and closely related taxonomic families (Figure 3). Thus, the Goldilocks Zone for RNA is not encompassed by the sampling regimes currently being employed for protein family discovery.

Discussion

Our analyses of over 400 publicly-available bacterial and archaeal RNA-seq datasets reveal that there is evidence for large numbers of RNAs of unknown function in public data. We find evidence for close to 1000 unannotated noncoding transcriptional outputs, but, given that RNA-seq experiments provide a snapshot of gene expression under specific experimental conditions, this number is likely to be far lower than the complete set of transcriptional outputs. Thus, the dataset we assembled for this project, which includes data generated by a number of labs and derives from various species and strains grown under a range of experimental conditions, is expected to represent a broad, though partial, census of total expression outputs across the species represented. Equally striking is the fact that, for the 922 RUFs identified in our study, over half (568) are among the most abundant transcripts. These results suggest that ncRNA may play an even greater role in the molecular workings of Bacteria and Archaea than hitherto realised.

This use of transcriptome data clearly improves our capacity to identify noncoding outputs: applying three criteria (sequence conservation, conservation of secondary structure, and expression in multiple strains/species) we have identified 163 high-confidence expressed RUFs from public data (Figure 2). An additional 405 RUFs are highly expressed across the transcriptomes we have examined, yet these do not show clear signs of sequence or structural conservation in other sequenced genomes. Given their high expression level, these seem unlikely to be transcriptional noise. Some may represent technical artefacts, but many could be *bona fide* lineage-specific ncRNAs with potentially novel functions.

Our results indicate that the greatest gain in analytical power for ncRNA discovery will come from phylogenetically-informed experimental design. Indeed, we find that this is critical to successful element identification, since the 'Goldilocks Zone' for optimal comparative analysis of RNA elements is surprisingly narrow. Hence, existing efforts to maximise phylogenetic coverage of genome space [1,2] need to be complemented with fine-scale sampling of the tips (Figure 4). Indeed, analysing the few transcriptomes that span the Goldilocks Zone reveals a remarkable enrichment of transcripts showing evidence of structure, conservation and expression in other strains/species. Furthermore, it is worth noting that the RNA family conservation decays as the phylogenetic distance increases (shown in Figure 3). There is a possibility that the Rfam families used for this are biased. However, if a bias exists, it is towards families with higher

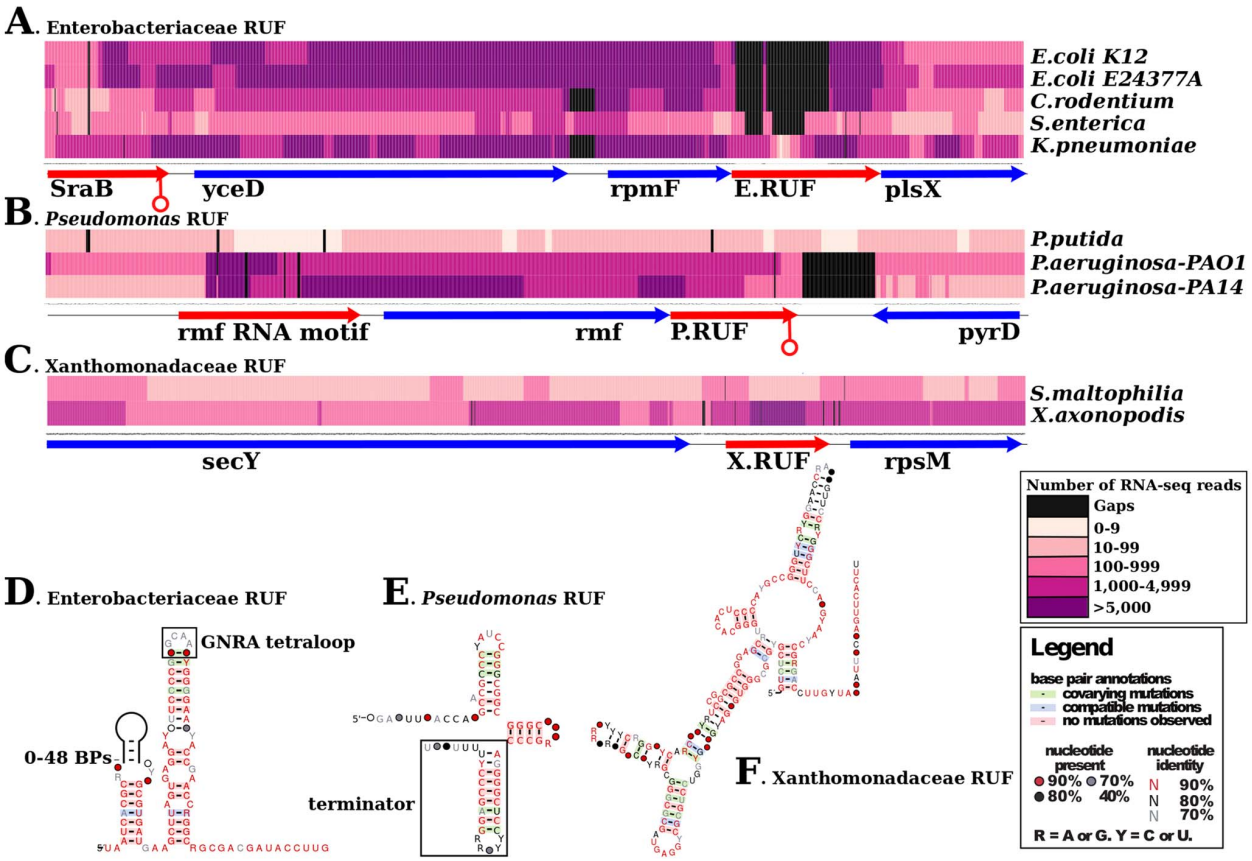

Figure 5. Comparative analysis of RNA-seq datasets in the Goldilocks Zone is a powerful approach for identifying RUFs. In this figure we illustrate data corresponding to 3 exemplar RUFs that show high covariation, conserved predicted secondary structures and are derived from one of the Goldilocks Zone clades shown in Figure 4. (**A–C**) The expression levels inferred from RNA-seq in the region encompassing each RUF. The regions contain a mix of ncRNAs (red arrows) and protein coding genes (blue arrows) and a RUF (red arrow). For each nucleotide, the total number of reads that map to that nucleotide was computed, and are presented as a heatmap; darker colours indicate high relative expression, lighter colours indicate low expression and black indicates a gap in the genomic alignment of the sequences for the locii. (**D–F**) R2R [68] representations of the predicted consensus secondary structures for exemplar RNAs of Unknown Function (RUFs) selected from the Enterobacteriaceae, *Pseudomonas* and Xanthomonadaceae data. Covariation is highlighted in green, structure-neutral variation is highlighted in blue, highly conserved regions are highlighted in pink. The Enterobacteriaceae RUF contains a conserved tetraloop of the GNRA or UNCG type, and there have been two independent insertions of hairpins in *S. enterica* and *K. pneumoniae* within the first hairpin. The *Pseudomonas* RUF hosts a 3′ rho independent transcription terminator.

doi:10.1371/journal.pcbi.1003907.g005

conservation (as the families are constructed from published ncRNAs that are often discovered based upon sequence conservation [22,35]). Thus, we might actually be overestimating RNA element conservation, making phylogenetically informed sampling even more important.

Given that isolation, cultivation and study of individual bacterial and archaeal strains can be extremely challenging [12] successful phylogeny-informed comparative RNA-seq will be a demanding endeavour, requiring complex sets of expertise spanning advanced culturing and isolation techniques, functional genomics capability and RNA bioinformatics. This places such a project beyond the reach of most individual labs. We therefore propose that comprehensive resolution of the comparative RNA-seq problem can best be resolved via a community-driven initiative: in recognition of the success of the GEBA project, we have dubbed this An RNA Encyclopedia of Bacteria and Archaea (AREBA). The appropriateness of this acronym will be especially clear to Japanophones, as, in Japanese, the phrase 'areba' (あれは) translates to 'if there'.

Materials and Methods

Preprocessing and mapping

All available bacterial and archaeal genomes were downloaded from the European Nucleotide Archive (ENA) (2,562 and 154 genomes, respectively) [50]. RNA-seq datasets published as of August 2013 were collected, spanning 37 species/strains, 44 experiments and 413 lanes of sequencing data (Table S2). Most of these datasets were generated on the Illumina platform [51], with a few lanes from the SOLiD platform [52] and the 454 platform [53]. Where possible, FastQ files were downloaded, scanned for residual adapter sequences using AdapterRemoval (v1.5.4) [54], and mapped to the reference genome using Bowtie2 (v2.1.0) [55] for Illumina and 454 data and BFAST (v0.7.0a) [56] for SOLiD data.

Producing consistent genome annotations

All genomes were re-annotated for both RNA genes and protein coding genes. Non-coding RNA genes were annotated using

cmsearch (v.1rc4) [57] to identify homologs of RNA families from the Rfam database (v11.0) using the default "gathering threshold" (cmsearch –cut_ga) [22,35]. Protein coding genes were annotated using three approaches: First, annotations were parsed from the ENA files. Secondly, Glimmer (v3.02) was run on all genomes to predict open reading frames (with parameters "-o7 -g45 -t15") [58]. Thirdly, all genomes were translated into all possible amino acid sequences of length 15 or more and scanned for homologs of entries in the Pfam database of protein families using hmmsearch (v3.1dev and the parameter "–cut_ga") [41,42].

Identification of novel RNAs

From the mapped RNA-seq data, potential novel RNA genes (designated RNAs of Unknown Function, or RUFs) were picked manually by locating regions in the genomes that showed high levels of expression without overlapping annotated protein coding or RNA genes. Only RUFs of lengths 50 to 400 nucleotides were included, yielding a total of 844 RUFs in Bacteria and 78 RUFs in Archaea.

Homology search and structure prediction

Homologs of the identified RUFs were found in all the downloaded genomes using nhmmer [59] in an iterative fashion: First, the RUF sequence alone was used in the scan; then, all hits with E-value <0.001 were included and a HMM built. This was iterated 5 times. The alignments from the RUF homology search were analyzed further by investigating the potential for secondary structure formation using RNAz [60] and alifoldz [61]. Protein coding potential of the RUFs was assessed using RNAcode [62]. Overlaps between potential RUF homologs in other strains/species and all the annotations in the respective genomes were also assessed.

Comparative expression and conservation analysis

For each strain, the available RNA-seq datasets were pooled and a list was created of transcripts showing expression in that strain in at least one experiment (defined as a transcript having a median depth of at least 10 reads in any experiment). A RUF homolog was defined as being expressed if the median read depth of the homologous region was at least 10X. Transcripts were ranked for each strain based on median expression (i.e. the most highly expressed transcript will have rank 1), which makes relative comparison across strains and datasets possible. The final set comprises 452 different Rfam families, 922 different RUFs, and 7249 different Pfam domains.

For comparative analysis, if a gene was found to be expressed in more than one strain/species, the minimum rank was used (i.e. showing the relatively most abundant expression of the gene). This ensures that transcripts that are always low abundance will remain low abundance, whereas genes that are highly abundant in at least one of the sampled time points and conditions will be treated as such. The ranking is used as a measure of expression.

"Family conservation" is based on SSU rRNA alignments of all Bacteria and Archaea, respectively. For each genome, the best hit to the Rfam model of SSU rRNA was extracted (RF00177 for Bacteria and RF01959 for Archaea). The sequences were aligned to the model using cmalign [57]. Finally, a distance matrix was calculated using dnadist [63] with the F84 model [64,65] which allows for different transition/transversion rates and for different nucleotide frequencies. The pairwise strain/species distances produced in this manner estimate the total branch length between any pair of strains/species. For any gene found in two or more strains/species, the maximum pairwise distance is used as the conservation score. Upper and lower quartiles of the distributions are used to define sets of high, medium and low expression and conservation, respectively. (Expression, upper quartile: 204. Expression, lower quartile: 1660. Conservation, upper quartile: 0.478. Conservation, lower quartile: 0.267).

Quality control of RNA-seq datasets

We ranked datasets based on the following quality control metrics (values reported in Table S3).

Strand correlation. We calculated correlation between the reads on the two strands. If the dataset is unstranded, we expect a correlation close to 1.

Expression of core genes. We defined a set of 40 core protein-coding genes based on [66,67] and 16 noncoding RNA genes (the union of tRNA, RNaseP, tmRNA, SRP, 6S and rRNA RNA families) [22,35]. If the median read depth is greater than 10X, we defined the gene as expressed. For each dataset, we report the fraction of the core genes that are expressed.

Coverage. We calculated coverage as the fraction of the genome covered by at least 10 mapped reads.

Fraction mapped reads. For each dataset, we ascertained the fraction of mapped reads.

Concordance. To measure how well a given RNA-seq dataset corresponds to the annotated genes in a genome, we developed a concordance metric. For this, we define true positives (TP) to be the number of annotated positions that are expressed; false positives (FP) to be the number of unannotated positions that are expressed; true negatives (TN) to be the number of unannotated positions that are not expressed; and false negatives (FN) to be the number of annotated positions that are not expressed. Note, not all annotated genes are expected to be expressed, and not all unannotated positions are false. Therefore, we calculate the positive predictive value (PPV):

$$PPV = \frac{TP}{TP+FP}$$

This measures the fraction of expressed positions that are annotated. We also calculate the fraction of the genome that is annotated:

$$ANN = \frac{TP+FN}{TP+FP+TN+FN}$$

To make the PPV more robust, our final concordance metric normalizes PPV by ANN.

Supporting Information

Table S1 The Pfam, Rfam and RUF identifiers for each entry corresponding to Figure 1.
(XLS)

Table S2 Strain/species names, genome accessions, RNA-seq data sources, Pubmed IDs, sequencing platform and notes for each dataset used for this study.
(XLS)

Table S3 Quality control measures computed for each RNA-seq dataset used in this study. The values are defined in detail in the Methods section.
(XLS)

Author Contributions

Conceived and designed the experiments: AMP PPG. Performed the experiments: SL SUU ASWL HE WL SM NEW LB PPG. Analyzed the data: SL SUU ASWL HE WL SM NEW LB PPG. Contributed reagents/materials/analysis tools: PJB NRT. Wrote the paper: AMP PPG SL SUU LB.

References

1. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462: 1056–60.
2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431–7.
3. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599–606.
4. Chun J, Rainey FA (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst Evol Microbiol 64: 316–24.
5. Sorek R, Serrano L (2011) Bacterial genomes: from regulatory complexity to engineering. Curr Opin Microbiol 14: 577–8.
6. Storz G, Vogel J, Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 43: 880–91.
7. Dennis PP, Omer A, Lowe T (2001) A guided tour: small RNA function in Archaea. Mol Microbiol 40: 509–19.
8. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327: 167–70.
9. Breaker RR (2012) Riboswitches and the RNA world. Cold Spring Harb Perspect Biol 4.
10. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157: 77–94.
11. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, et al. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571–9.
12. Stewart EJ (2012) Growing unculturable bacteria. J Bacteriol 194: 4151–60.
13. Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, et al. (2008) A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci U S A 105: 8102–7.
14. Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543–6.
15. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.
16. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443: 950–5.
17. Mandin P, Toledo-Arana A, Fouquier d'Hérouel A, Repoila F (2013) RNA-mediated control of bacterial gene expression: role of regulatory non-coding RNAs. Encyclopedia of Molecular Cell Biology and Molecular Medicine. Wiley-VCH Verlag GmbH & Co. KGaA, pp.1–36.
18. Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17: 117–125.
19. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–7.
20. Barrick JE, Breaker RR (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol 8: R239.
21. von Hippel PH (1998) An integrated model of the transcription complex in elongation, termination, and editing. Science 281: 660–5.
22. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 39: D141–5.
23. Santangelo TJ, Artsimovitch I (2011) Termination and antitermination: RNA polymerase runs a stop sign. Nat Rev Microbiol 9: 319–29.
24. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, et al. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell 31: 147–57.
25. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S (1983) The RNA moiety of Ribonuclease P is the catalytic subunit of the enzyme. Cell 35: 849–57.
26. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428: 281–6.
27. Roth A, Weinberg Z, Chen AG, Kim PB, Ames TD, et al. (2014) A widespread self-cleaving ribozyme class is revealed by bioinformatics. Nat Chem Biol 10: 56–60.
28. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709–12.
29. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960–4.
30. Narberhaus F, Waldminghaus T, Chowdhury S (2006) RNA thermometers. FEMS Microbiol Rev 30: 3–16.
31. Loh E, Kugelberg E, Tracy A, Zhang Q, Gollan B, et al. (2013) Temperature triggers immune evasion by Neisseria meningitidis. Nature 502: 237–40.
32. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, et al. (2000) Homologs of small nucleolar RNAs in Archaea. Science 288: 517–22.
33. Gaspin C, Cavaillé J, Erauso G, Bachellerie JP (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. J Mol Biol 297: 895–906.
34. Gardner PP, Bateman A, Poole AM (2010) SnoPatrol: how many snoRNA genes are there? J Biol 9: 4.
35. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, et al. (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 41: D226–32.
36. Hoeppner MP, Gardner PP, Poole AM (2012) Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Comput Biol 8: e1002752.
37. Croucher NJ, Thomson NR (2010) Studying bacterial transcriptomes using RNA-seq. Curr Opin Microbiol 13: 619–24.
38. van Opijnen T, Camilli A (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11: 435–42.
39. Barquist L, Boinett CJ, Cain AK (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. RNA Biol 10: 1161–9.
40. Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, et al. (2013) A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. Nucleic Acids Res 41: 4549–64.
41. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290–301.
42. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222–30.
43. McCutcheon JP, Eddy SR (2003) Computational identification of non-coding rnas in saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res 31: 4119–28.
44. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–8.
45. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927–31.
46. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338–43.
47. Gottesman S, Storz G (2011) Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol 3: a003798
48. Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biol 3: e10.
49. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33: 2433–9.
50. Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A, et al. (2013) Facing growth in the European Nucleotide Archive. Nucleic Acids Res 41: D30–5.
51. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728–32.
52. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613–9.
53. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–80.
54. Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5: 337.
55. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–9.
56. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4: e7767.
57. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–5.
58. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673–9.
59. Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. Bioinformatics 29: 2487–9.
60. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF (2010) RNAz 2.0: improved noncoding RNA detection. Pac Symp Biocomput 2010: 69–79.
61. Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 342: 19–30.
62. Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, et al. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17: 578–94.
63. Felsenstein J (2005) Phylip (phylogeny inference package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

64. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29: 170–9.
65. Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93–104.
66. Wu D, Jospin G, Eisen JA (2013) Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS One 8: e77033.
67. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, et al. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2: e243.
68. Weinberg Z, Breaker RR (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics 12: 3.

CHAPTER III - A Benchmark of RNA-RNA Interactions

Many ncRNAs utilize RNA-RNA interactions. The list of such RNAs spanning through all domains of life, including: bacterial and archaeal sRNAs (Storz et al. 2011; Prasse et al. 2013; Babski et al. 2014), siRNAs (Carthew & Sontheimer 2009; Borges & Martienssen 2015), miRNAs (Carthew & Sontheimer 2009; Cuperus et al. 2011; Chen 2008), snRNAs (Karijolich & Yu 2010), snoRNAs (Brown et al. 2001; Kiss 2002; Gardner et al. 2010; Omer et al. 2000), scaRNAs (Darzacq et al. 2002), CRISPR RNAs (Bhaya et al. 2011; Barrangou et al. 2007) and piRNAs (Klattenhoff & Theurkauf 2008; Brennecke et al. 2007).

With ncRNA functions often being mediated through RNA-RNA interactions, there has been considerable focus on developing algorithms and software that can predict interactions. The aim of the following study is to benchmark available RNA-RNA interaction prediction tools as a way of assessing their performance and utility. To achieve this goal, we compiled a dataset of verified RNA interactions and assessed performance of 15 tools. Our results show that the energy-based methods with accessibility are the most successful programs which is similar with the current literature (Pain et al. 2015; Lai & Meyer 2015). The accessibility methods calculate the energy needed to open designated binding regions of interacting RNAs (Richter & Backofen 2012; Lorenz et al. 2011; Tafer & Hofacker 2008), which is considered biophysically the most sensible binding model (Richter & Backofen 2012) (Figure 3.1). Our findings informed our decision to use the RNAup algorithm (Mückstein et al. 2006) (an accessibility method) for predicting RNA-RNA interactions in the work presented in Chapter IV, because RNAup produces better overall scores than any other methods.

We submitted this work to the journal *Bioinformatics*, and the manuscript, as submitted, is presented below.

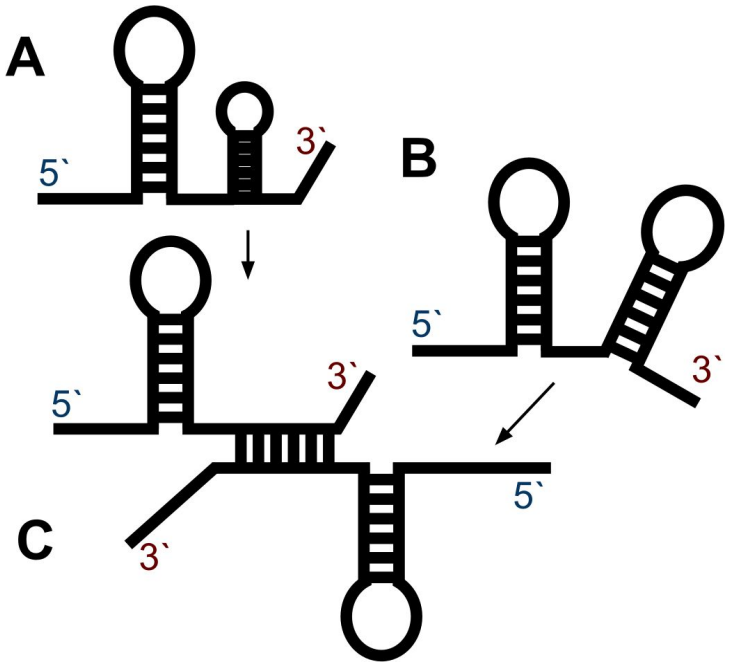

Figure 3.1 An example duplex structure of two interacting RNAs (seen in C). An accessibility based energy model includes the opening energy penalty required to make binding regions available. If the internal structure of interacting RNAs are too stable at binding regions, this will lead to thermodynamically infeasible (unstable) structures with high duplex MFE. When two RNAs are in their native shape without any interaction, they contain internal base-pairings (seen in A and B). On the other hand, if the designated binding regions are accessible, this will create a stable duplex structure with low MFE (seen in C).

REFERENCES

Babski, J. et al., 2014. Small regulatory RNAs in Archaea. *RNA biology*, 11(5), pp.484–493.

Barrangou, R. et al., 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), pp.1709–1712.

Bhaya, D., Davison, M. & Barrangou, R., 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, pp.273–297.

Borges, F. & Martienssen, R.A., 2015. The expanding world of small RNAs in plants. *Nature*

reviews. Molecular cell biology, 16(12), pp.727–741.

Brennecke, J. et al., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6), pp.1089–1103.

Brown, J.W. et al., 2001. Multiple snoRNA gene clusters from Arabidopsis. *RNA* , 7(12), pp.1817–1832.

Carthew, R.W. & Sontheimer, E.J., 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4), pp.642–655.

Chen, X., 2008. MicroRNA metabolism in plants. *Current topics in microbiology and immunology*, 320, pp.117–136.

Cuperus, J.T., Fahlgren, N. & Carrington, J.C., 2011. Evolution and functional diversification of MIRNA genes. *The Plant cell*, 23(2), pp.431–442.

Darzacq, X. et al., 2002. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *The EMBO journal*, 21(11), pp.2746–2756.

Gardner, P.P., Bateman, A. & Poole, A.M., 2010. SnoPatrol: how many snoRNA genes are there? *Journal of biology*, 9(1), p.4.

Karijolich, J. & Yu, Y.-T., 2010. Spliceosomal snRNA modifications and their function. *RNA biology*, 7(2), pp.192–204.

Kiss, T., 2002. Small Nucleolar RNAs. *Cell*, 109(2), pp.145–148.

Klattenhoff, C. & Theurkauf, W., 2008. Biogenesis and germline functions of piRNAs. *Development* , 135(1), pp.3–9.

Lai, D. & Meyer, I.M., 2015. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic acids research*, 44(7), p.e61.

Lorenz, R. et al., 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology: AMB*, 6, p.26.

Mückstein, U. et al., 2006. Thermodynamics of RNA–RNA binding. *Bioinformatics* , 22(10), pp.1177–1182.

Omer, A.D. et al., 2000. Homologs of small nucleolar RNAs in Archaea. *Science*, 288(5465), pp.517–522.

Pain, A. et al., 2015. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5), pp.509–513.

Prasse, D. et al., 2013. Regulatory RNAs in archaea: first target identification in

Methanoarchaea. *Biochemical Society transactions*, 41(1), pp.344–349.

Richter, A.S. & Backofen, R., 2012. Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA biology*, 9(7), pp.954–965.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Tafer, H. & Hofacker, I.L., 2008. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* , 24(22), pp.2657–2663.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Structural bioinformatics

A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life

Sinan Uğur Umu [1,2,*], and Paul P. Gardner [1,2,3]

[1]School of Biological Sciences, [2]Biomolecular Interaction Centre and [3]Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The aim of this study is to assess the performance of RNA-RNA interaction prediction tools for all domains of life.

Results: Minimum free energy (MFE) and alignment methods constitute most of the current RNA interaction prediction algorithms. The MFE tools that include accessibility (i.e. RNAup, IntaRNA and RNAplex) to the final predicted binding energy have better true positive rates (TPRs) with a high positive predictive values (PPVs) in all datasets than other methods. They can also differentiate almost half of the native interactions from background. The algorithms that include effects of internal binding energies to their model and alignment methods seem to have high TPR but relatively low associated PPV compared to accessibility based methods.

Availability: We shared our wrapper scripts and datasets at Github (github.com/UCanCompBio/RNA_Interactions_Benchmark). All parameters are documented for personal use.

Contact: sinan.umu@pg.canterbury.ac.nz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA biology has become more prominent after the discovery of non-coding RNAs (ncRNAs) and their versatile functions (Mattick, 2004; Ambros, 2004; Kidner and Martienssen, 2005; Mattick, 2009; Waters and Storz, 2009; Storz *et al.*, 2011; Barquist and Vogel, 2015). The versatility of RNA molecules has led to the idea of an "RNA world" where RNA formed the first primitive life forms (Gilbert, 1986). The importance of RNA biology is highlighted by the relatively small fraction of protein-coding regions of most eukaryotic genomes (Mattick, 2004, 2009). For example, 1.2% of the human genome contain protein coding genes, while 76% is transcribed into RNA (Pennisi, 2012). Likewise, prokaryotic cells contain various ncRNAs genes (Vogel, 2009; Holmqvist and Vogel, 2013; Gottesman, 2004; Thébault *et al.*, 2014) and have also been shown to have transcriptional complexity like eukaryotes (Güell *et al.*, 2011, 2009; Barquist and Vogel, 2015; Lindgreen *et al.*, 2014; Cohen *et al.*, 2016).

ncRNA molecules often utilize RNA-RNA base pairing like bacterial/archaeal small RNAs (sRNAs) (Storz *et al.*, 2011; Prasse *et al.*, 2013), small interfering RNAs (siRNAs) (Carthew and Sontheimer, 2009), microRNAs (miRNAs) (Carthew and Sontheimer, 2009; Cuperus *et al.*, 2011), spliceosomal small nuclear RNAs (snRNAs) (Karijolich and Yu, 2010), small nucleolar RNAs (snoRNAs) (Brown *et al.*, 2001; Kiss, 2002; Gardner *et al.*, 2010; Omer *et al.*, 2000), cajal-body specific small nuclear RNAs (scaRNAs) (Darzacq *et al.*, 2002), clustered regularly-interspaced short palindromic repeats (CRISPR) RNA (Bhaya *et al.*, 2011) and piwi-interacting RNAs (piRNAs) (Klattenhoff and Theurkauf, 2008; Brennecke *et al.*, 2007). It seems some long-noncoding RNAs (lncRNAs) may also engage RNA-RNA interactions (Kung *et al.*, 2013), which are quite abundant in eukaryotes (Zhao *et al.*, 2016).

In addition to endogenous ncRNAs genes, many experimental techniques take advantage of RNA-RNA interactions such as gene silencing (i.e. knock-out) by artificial siRNAs (Deleavey and Damha, 2012; Reynolds *et al.*, 2004) and designing oligonucleotides for ribosomal RNA (rRNA) depletion in RNA-seq experiments (O'Neil *et al.*, 2013).

Different clades of life utilize regulatory RNA-RNA interactions with different constraints: various mediator proteins (Vogel and Luisi, 2011; Carthew and Sontheimer, 2009), binding regions preference and distinct complementarity requirements (Millar and Waterhouse, 2005; Ameres and Zamore, 2013). Thus, many different tools have been developed to predict stable interactions. Some algorithms solve RNA-RNA interaction as an alignment problem using local alignment approaches (Wenzel *et al.*,

2012; Hodas and Aalberts, 2004). Most of these use dynamic programming and minimum free energy methods (MFE) (Dieterich and Stadler, 2012; Lorenz *et al.*, 2011; Backofen and Hess, 2010), which are also widely used methods for RNA secondary structure predictions (Zuker and Sankoff, 1984; Zuker and Stiegler, 1981; McCaskill, 1990; Nussinov and Jacobson, 1980; Zuker, 2000; Markham and Zuker, 2008). In bacteria, comparative methods are becoming popular (Wright *et al.*, 2013; Kery *et al.*, 2014; Pain *et al.*, 2015), but they are restricted to conserved sRNAs, which are quite rare (Lindgreen *et al.*, 2014; Barquist and Vogel, 2015).

RNA target detection is still a challenging task but it is vital to understand more about RNA-RNA interactions for functional annotation of unknown transcripts while making computationally feasible and biologically relevant predictions. In this study, we assessed the performance of available RNA-RNA interaction prediction tools on trusted, verified datasets from all domains of life. We evaluated their ability to find true RNA-RNA pairs by calculating TPRs, PPVs and Matthews correlation coefficients (MCCs) (Matthews, 1975) in eukaryotic, bacterial and archaeal systems. We also assessed how successfully they predict binding scores and reported the significance of these predictions.

2 Materials and Methods

All RNA interaction prediction algorithms are freely available and cited in the manuscript. We used Python, R, Bash for the scripts and wrappers, which are shared in our Github repository (github.com/UCanCompBio /RNA_Interactions_Benchmark). A parser script (or a wrapper script) has been written for each of the tools benchmarked here. All the parameters and command line arguments are also accessible.

2.1 Benchmark datasets

We manually confirmed the correct interaction regions for all dataset items and used entire target regions (i.e. UTRs, coding regions or target RNA) without any truncation to make our benchmark as realistic as possible.

The eukaryotic benchmark dataset consisted of miRNAs from human, *Arabidopsis, Caenorhabditis elegans (C. elegans)* (Kozomara and Griffiths-Jones, 2013; Chou *et al.*, 2015); C/D and H/ACA box snoRNAs from human, *Arabidopsis, C. elegans*, yeast (Brown *et al.*, 2001; Yoshihama *et al.*, 2013; Lestrade and Weber, 2006; Piekna-Przybylska *et al.*, 2007); human and yeast U6/U2 snRNAs (Will and Lührmann, 2011); endogenous siRNAs from *Arabidopsis* (Addo-Quaye *et al.*, 2008) and piRNAs from mouse (Gou *et al.*, 2015). Experimentally verified miRNA/siRNA/piRNA-target mRNAs and snoRNA/snRNA-target RNAs were selected from different ncRNA families as much as possible (in total 88 pairs) (Supplementary Table S1).

We compiled a bacterial sRNA and target mRNA dataset from *Salmonella, Escherichia coli (E. coli)* and *Listeria monocytogenes (L. monocytogenes)* that consists of 60 verified sRNA-mRNA pairs (Cao *et al.*, 2010; Peer and Margalit, 2011; Lai and Meyer, 2015). The target regions of bacterial sRNAs lie either in 5‘UTR or downstream of start codon (Storz *et al.*, 2011; Richter and Backofen, 2012). We selected regions 200 nucleotides (nts) upstream to 100 nts downstream of the start codons (i.e. 5‘end mRNA) which contain verified binding regions. We extracted both sRNAs and target 5‘ end mRNAs from their associated genome sequences (Acces. AE006468.1, AL591824.1 and U00096.3) (Supplementary Table S1).

We gathered a set of archaeal C/D box snoRNAs consisting of 5 snoRNAs and their ribosomal RNA targets (Omer *et al.*, 2000). We also added a member of less studied archaeal sRNA (from *Methanosarcina mazei*) (Jäger *et al.*, 2012). Selected genes and targets were obtained from their associated archaeal genomes (AE008384.1) or Genbank (Supplementary Table S1).

2.2 Accuracy measures

We calculated TPR (sensitivity) and PPV (precision) scores of each algorithm based on prediction of binding region for all 154 interactions. Therefore, true positives (TPs) are the number of correctly predicted pairings (of the RNA duplex), false positive (FPs) are the number of falsely predicted base pairings, and false negatives (FNs) are the missed base pairings on the targets. True negatives (TNs) are not applicable to our predictions, as TN numbers increase with the size of target RNAs and are bounded by the total number of predicted base-pairs. However, we calculated approximated MCCs (Matthews, 1975) by using the geometric mean of TPR and PPV (Wenzel *et al.*, 2012; Gorodkin *et al.*, 2001). These can be defined as:

$$TPR(sensitivity) = TP/(TP + FN) \quad (1)$$

$$PPV(precision) = TP/(TP + FP) \quad (2)$$

$$MCC \approx \sqrt{TPR \cdot PPV} \quad (3)$$

Besides the well known accuracy measures, we also assessed the scores generated by the algorithms, which usually show the stability of interaction (e.g. a binding MFE). For each true and verified target (positive control), we created 200 dinucleotide shuffled sequences (negative controls) using the esl-shuffle tool (Eddy, 2011) to prevent possible biases caused by the nearest-neighbour energy model of structure prediction (Workman and Krogh, 1999). As a further test to determine the significance of native interactions, we fitted shuffled interactions (as a background) into both normal and Gumbel distributions (using negative energies) (Gumbel, 1958), since MFE values mostly follow an extreme value distribution (Rehmsmeier *et al.*, 2004; Tjaden, 2008). We applied this approach only to bacterial dataset due to time constraints and the uniform distribution of bacterial targets (i.e. targets 300 nts long).

We selected the best scoring interaction as the native interaction if an algorithm produces more than one interaction, which is also true for all our analyses.

3 Results and Discussion

3.1 RNA-RNA interaction prediction tools

The RNA-RNA interaction prediction methods are divided mainly into three groups: alignment like methods, MFE methods and comparative (homology) methods. We can also further divide the MFE methods into three different sub-classes based on whether their approach considers intramolecular base-pairs (internal structure), neglects intramolecular structure or measures the accessibility of the binding region. There are also other machine learning algorithms (Oğul *et al.*, 2011; Yang *et al.*, 2008), and probabilistic approaches like RactIP (Kato *et al.*, 2010), which uses the CONTRAfold model (Do *et al.*, 2006) for RNA interaction prediction.

RIsearch (Wenzel *et al.*, 2012), Bindigo (Hodas and Aalberts, 2004) and Guugle (Gerlach and Giegerich, 2006) are examples of alignment-like methods. The RIsearch algorithm was mainly developed for rapidly searching genomes to detect RNA-RNA pairs from genome sequencing data by combining the Smith-Waterman-Gotoh algorithm with a nearest-neighbor energy model (Wenzel *et al.*, 2012), while Bindigo adopts an optimized Smith-Waterman to find optimal oligonucleotide-RNA pairs (Hodas and Aalberts, 2004). Guugle uses suffix arrays to seek RNA targets based on RNA helix rules that allow G-U pairs (Gerlach and Giegerich, 2006).

Besides these alignment based methods, tools like BLAST (Altschul *et al.*, 1990), Blat (Kent, 2002), ssearch (Pearson and Lipman, 1988) or other local alignment implementations can be used to rapidly collect long

(reverse) complementary regions by including G-U pairs (C-U or G-A for the reverse complement) in the scoring matrix (Gerlach and Giegerich, 2006; Wenzel *et al.*, 2012; Thébault *et al.*, 2014).

MFE methods form the majority of the RNA-RNA interaction prediction tools (Dieterich and Stadler, 2012; Lorenz *et al.*, 2011; Backofen and Hess, 2010). Many secondary structure prediction tools also utilize MFE methods (Mathews and Turner, 2006; Lorenz *et al.*, 2011; Zuker and Sankoff, 1984; Markham and Zuker, 2008). Some MFE methods including RNAhybrid (Rehmsmeier *et al.*, 2004), RNAduplex (Lorenz *et al.*, 2011), DuplexFold (Reuter and Mathews, 2010) and TargetRNA (Tjaden, 2008) neglect intramolecular structures for the sake of algorithmic speed. Algorithms like Pairfold (Andronescu *et al.*, 2005), RNAcofold (Bernhart *et al.*, 2006) and bifold (Reuter and Mathews, 2010) take intramolecular base-pairing into account. RNAup (Mückstein *et al.*, 2006), RNAplex (Tafer and Hofacker, 2008) and IntaRNA (Busch *et al.*, 2008) compute the accessibility of binding regions to report the final MFE of the RNA duplex, which is considered more realistic biophysically (Richter and Backofen, 2012). AccessFold includes accessibility using a method defined as pseudo-energy minimization (DiChiacchio *et al.*, 2015). BistaRNA also includes accessibility and can predict multiple binding sites (Poolsap *et al.*, 2011). Lastly, tools like TargetRNA2 (Kery *et al.*, 2014), CopraRNA (Wright *et al.*, 2013), miRanda (John *et al.*, 2004), TargetScan (Lewis *et al.*, 2005), PETcofold (Seemann *et al.*, 2011) and DIANA-microT (Kiriakidou *et al.*, 2004) exploit homology and evolutionary conservation to predict interactions

Some RNA-RNA interaction prediction tools are developed to achieve a specific task or to predict very specific group of interactions. For example, PLEXY is designed for C/D snoRNAs (Kehr *et al.*, 2011), RNAsnoop (Tafer *et al.*, 2010) for H/ACA snoRNAs and TargetRNA (Tjaden, 2008) for bacterial sRNAs (*E. coli* and *Salmonella*). In this study, we tried to assess the versatility of prediction tools on different datasets as well as their prediction power where applicable. We excluded tools designed for specific RNA families such as specialized miRNA algorithms (reviewed in Witkos *et al.* 2011), specialized snoRNA target prediction algorithms and comparative bacterial sRNA prediction methods (reviewed in Backofen and Hess 2010, Pain *et al.* 2015). We also excluded inteRNA (Alkan *et al.*, 2006), IRIS (Pervouchine, 2004), piRNA (Chitsaz *et al.*, 2009b) and biRNA (Chitsaz *et al.*, 2009a), as they are either no longer supported or obsolete.

In summary, our final list of selected tools used for further analyses consisted of RIsearch (Wenzel *et al.*, 2012), IntaRNA (Busch *et al.*, 2008), RNAcofold (Bernhart *et al.*, 2006), RNAhybrid (Rehmsmeier *et al.*, 2004), RNAduplex (Lorenz *et al.*, 2011), RNAplex (Tafer and Hofacker, 2008), RNAup (Mückstein *et al.*, 2006), pairfold (Andronescu *et al.*, 2005), bifold (Reuter and Mathews, 2010), DuplexFold (Reuter and Mathews, 2010), ssearch (Pearson, 1991), RactIP (Kato *et al.*, 2010), bistaRNA (Poolsap *et al.*, 2011), AccessFold (DiChiacchio *et al.*, 2015) and NUPACK (Dirks *et al.*, 2007) (Supplementary Table S2).

3.2 Overall prediction performances

Our analyses of the overall performances of RNA interaction prediction algorithms show that three accessibility based algorithms (RNAup, IntaRNA and RNAplex) scored highest for sensitivity and precision. RNAup was highly precise compared to other tools (Figure 1 and Table 1). IntaRNA was the second algorithm (almost identical to RNAup) with a reasonable running time. RNAplex was comparable to both algorithms. RNAduplex had the best overall TPR score, but it was not as precise as IntaRNA. Table 1 summarizes the 'cumulative' TPR, PPV and MCC scores, while Figure 1 shows their distribution for all interactions (n=154) on all domains of life.

RIsearch and ssearch were the fastest methods, but they were not very sensitive or precise (Table 1). AccessFold and bifold had the longest run time, which appeared to increase for long RNA sequences like ribosomal RNAs or large target UTR regions. RIsearch and bifold gave inconsistent results, with combined MCCs of 0.33 and 0.40 respectively (Table 1). However, if we use a distribution of results as in Figure 1, the median MCCs appear to be zero for these algorithms. As bifold frequently returned no duplex structures for some RNA pairs (e.g. *C. elegans* miRNAs lin-4, lsy-6-3p etc.), and RIsearch produced many unsuccessful predictions for bacterial sRNAs, which produced to zero MCC scores for both.

Table 1. Total run time of algorithms, and the cumulative TPR, PPV and MCC scores.

Algorithm	Total run time (s) on selected files (n=50)	TPR (Sensitivity)	PPV (Precision)	MCC
AccessFold	596.44	0.38	0.31	0.35
bifold	404.63	0.37	0.31	0.34
bistaRNA	102.29	0.15	0.16	0.15
DuplexFold	**5.33**	0.48	0.17	0.29
IntaRNA	24.44	**0.59**	**0.56**	**0.58**
NUPACK	794.2	0.42	0.42	0.42
pairfold	90.24	0.39	0.29	0.34
ractIP	87.62	0.16	0.06	0.1
RIsearch	**4.16**	0.36	0.45	0.40
RNAcofold	15.28	0.41	0.32	0.36
RNAduplex	6.45	**0.66**	0.12	0.27
RNAhybrid	32.84	**0.56**	0.12	0.26
RNAplex	17.19	0.55	**0.57**	**0.56**
RNAup	137.48	0.51	**0.69**	**0.60**
ssearch	**4.69**	**0.56**	0.1	0.23

The cumulative scores (i.e. TPR, PPV, MCC) are calculated by adding individual TP, FP and FN values for all predictions.

3.3 The significance test results of bacterial dataset

The MFE values produced by the algorithms are not very explicit, so it is common to use negative controls to determine the significance of predicted energy values (Rehmsmeier *et al.*, 2004), especially for structure predictions (Workman and Krogh, 1999). We created negative controls for each pair as explained in materials and methods. Some algorithms were excluded from this assessment, because either they do not produce a score (i.e. RactIP, bistaRNA and ssearch) or are biased towards internal structures (i.e. pairfold, RNAcofold, bifold and NUPACK). Thus, the test of significance includes only 8 prediction algorithms (Table 2).

Table 2. The test of significance results of selected algorithms on bacterial sRNAs.

Algorithm	Total # of significant correct predictions for Gumbel dist. (n=60)	Total # of significant correct predictions for normal dist. (n=60)	Median rank of native interactions
AccessFold	15	17	41.75
DuplexFold	2	8	63.5
IntaRNA	**23**	**26**	**19**
RIsearch	13	14	52.25
RNAduplex	8	11	54.25
RNAhybrid	5	6	76
RNAplex	**23**	**30**	**10.5**
RNAup	**28**	**29**	**13.5**

Higher is better for the second and third columns. Lower is better for the fourth column.

These results show that RNAplex and RNAup reported almost half of the native energies as significant if they are fitted to normal distributions. It seems the Gumbel fitting of scores is more conservative which likely decreases the risk of FP predictions on high-throughput predictions. RNAup results were almost identical for both distributions. IntaRNA performed slightly worse than these two algorithms. The last column of Table 2 shows the median rank of native interactions. If a prediction score of a native interaction has the highest score (e.g. lowest MFE), it is ranked 1 out of 201. Therefore, the median ranks in the last column can be interpreted as the expected number of FPs introduced by the algorithms before predicting the native interaction.

3.4 A summary of RNA-RNA interactions and algorithm performances for all domains of life

Eukaryotic RNA interactions mostly focus on RNA interference (RNAi) (i.e. miRNAs and siRNAs) (Carthew and Sontheimer, 2009; Ambros, 2004; Chen, 2008). In animal RNAi, miRNAs (~20 nts long) prefer perfect complementarity in the seed region and have overall lower complementarity than plant counterparts (Axtell *et al.*, 2011; Ameres and Zamore, 2013). In plants, high complementary target regions may lie in coding region as well as UTRs rather than only 3'UTRs (Millar and Waterhouse, 2005; Axtell *et al.*, 2011; Ameres and Zamore, 2013). It is possible for a miRNA to target more than one region, especially in animals, which is known to increase efficiency of target gene downregulation (Millar and Waterhouse, 2005). However, in our benchmark we preferred to select miRNA targets containing a single designated binding region. Piwi associated piRNAs are also small endogenous RNAs (24-30 nts long) (Klattenhoff and Theurkauf, 2008; Zhang *et al.*, 2015), some of which use antisense binding to regulate target RNAs (Gou *et al.*, 2015) like miRNA and siRNA. H/ACA and C/D snoRNAs have roles in rRNA and snRNA maturation (Brown *et al.*, 2001; Kiss, 2002; Gardner *et al.*, 2010). These interactions differ in that C/D snoRNAs prefer a binding region on target RNA with consecutive nts around 7 to 20 bases long with a few mismatches (Gardner *et al.*, 2010; Kehr *et al.*, 2011), while H/ACA snoRNAs contain a stem loop within the binding region, which complicates target prediction (Kiss *et al.*, 2004; Tafer *et al.*, 2010; Gardner *et al.*, 2010). Spliceosomal snRNAs form ribonucleoprotein (RNP) complexes with other snRNAs (Karijolich and Yu, 2010), and they are also targeted by snoRNAs (termed scaRNAs) (Darzacq *et al.*, 2002). We included examples of both snRNA-snRNA and scaRNA-snRNA interactions to our dataset. It is also known that some lncRNAs use RNA-RNA interactions (Kung *et al.*, 2013) but these were not included in our benchmark.

We found that in the eukaryotic dataset, accessibility based methods performed best based on the average MCC scores (except AccessFold and bistaRNA) (Figure 2). IntaRNA (av. MCC: 0.51) slightly outperformed RNAup (av. MCC: 0.49) and produced a higher PPV than the other tools benchmarked. RNAplex (av. MCC: 0.48) and RIsearch (av. MCC: 0.48) (an alignment-like method) were also comparable with these two algorithms for eukaryotic datasets. Supplementary Table S3 explicitly shows the prediction scores for all 88 eukaryotic interactions.

Bacterial small RNAs can be divided into three major types: antisense binding sRNAs, *Hfq* dependent sRNAs and csrA binding sRNAs (Vogel, 2009; Storz *et al.*, 2011). However, in this study, bacterial sRNAs refer to either antisense or *Hfq* dependent sRNAs, which achieve their role via RNA-RNA base-pairing interactions. Bacterial sRNAs (50-200 nts long) prefer short binding regions relative to their size (Vogel, 2009; Storz *et al.*, 2011). This was also true for our dataset, with an average binding region size of 23 nts, with the smallest just 7 nts long (Supplementary Table S1). Model bacterial organisms like *E. coli* or *Salmonella* contain hundreds of different sRNAs which points to a complex regulatory system in prokaryotic organisms (Waters and Storz, 2009). Moreover, increasing number of RNA-seq studies (Sharma and Vogel, 2014; Sharma *et al.*, 2010; Cohen *et al.*, 2016) reveal that there are novel regulatory ncRNAs are spanning in prokaryotes than previously anticipated (Lindgreen *et al.*, 2014; Barquist and Vogel, 2015; Chen *et al.*, 2016).

We found that in the bacterial dataset, accessibility based methods performed better than the others based on the average MCC scores, as with the eukaryotic dataset. RNAup (av. MCC: 0.68) slightly outperformed IntaRNA (av. MCC: 0.65) in bacterial sRNA interactions. RNAplex (av. MCC: 0.61) was comparable with the other two algorithms. In bacterial dataset, RIsearch (av. MCC: 0.31) did not perform as well as on the eukaryotic dataset, which decreased the overall performance (Figure 2).

RNA interactions in archaea are not well characterized. Recent studies have shown that archaeal genomes contain a large number of ncRNA repositories similar to bacterial genomes (Lindgreen *et al.*, 2014). Unfortunately, there are not many verified RNA interactions available in archaea, except archaeal snoRNAs. Archaeal genomes mostly contain C/D box snoRNAs; thus, we added 5 C/D box snoRNAs (Omer *et al.*, 2000) and one archaeal sRNA (Jäger *et al.*, 2012) as an archaeal benchmark dataset. The archaeal sRNA targets a bicistronic gene and trans-regulates expression of two protein coding genes concurrently (Jäger *et al.*, 2012) (Figure 1, Figure 2 and Supplementary Table S3).

We found that in the archaeal dataset, RNAplex (av. 0.65) performed better than the other algorithms, followed by IntaRNA (av. MCC: 0.61). These two algorithms were followed by RNAup (av. MCC: 0.53) and RIsearch (av. MCC: 0.40). RIsearch was better on snoRNA predictions than the single archeal sRNA, which reduced the average overall performance. RNAplex recovered the binding region with a perfect MCC score, followed by IntaRNA.

3.5 Limitations of RNA-RNA interaction predictions algorithms

Unfortunately, 15 out of 154 RNA interaction pairs in our benchmark dataset could not be correctly predicted by any of the algorithms (i.e. an MCC score of 0 for all algorithms) (Figure 2 and Supplementary Table S3) including 6 human miRNAs, and snoRNAs from yeast, human and archaea. The mouse piRNA results were also unsatisfactory, and one (piR-013474) could not be detected by any of the algorithms. The algorithms benchmarked performed best on *Arabidopsis* miRNAs, siRNAs and bacterial sRNAs (Figure 2).

We applied the significance test to some of these failed eukaryotic interactions (e.g. mouse piRNAs, human miRNAs), aiming to see whether the predicted scores enabled the detection of true interactions (and separate scores for native interactions from background) rather than using correctly predicted base-pairs, which were used to calculate TPs. The comparison of two methods revealed consistent results as expected. For example, the native interaction of piR-013474 cannot be differentiated from background by any algorithm. This is also similar for other piRNAs and human miRNAs, where all algorithms consistently failed.

The lengths of target RNA regions (which include binding regions) seem to influence prediction quality (also discussed by Lai and Meyer 2015). The average length of a eukaryotic target RNA is 1690 nts long in our dataset. However, this rises to around 2400 nts for those miRNAs which did not give prediction scores, and longer in piRNAs. As described in materials and methods, we did not truncate the targets (e.g. UTRs) that contained binding regions. We found a significant reverse correlation (Pearson's r = - 0.28, $p < 0.05$) between the lengths of target RNAs and average MCCs (i.e. overall performances). However, some of the algorithms (RNAup, RNAplex, RIsearch, RNAcofold and NUPACK) are less prone to this length bias ($p > 0.05$) (Supplementary Table 4), making them ideal for use on untruncated targets.

Another explanation for inadequate prediction may be the quality of the dataset. Not all experimental protocols are equally strong at detecting

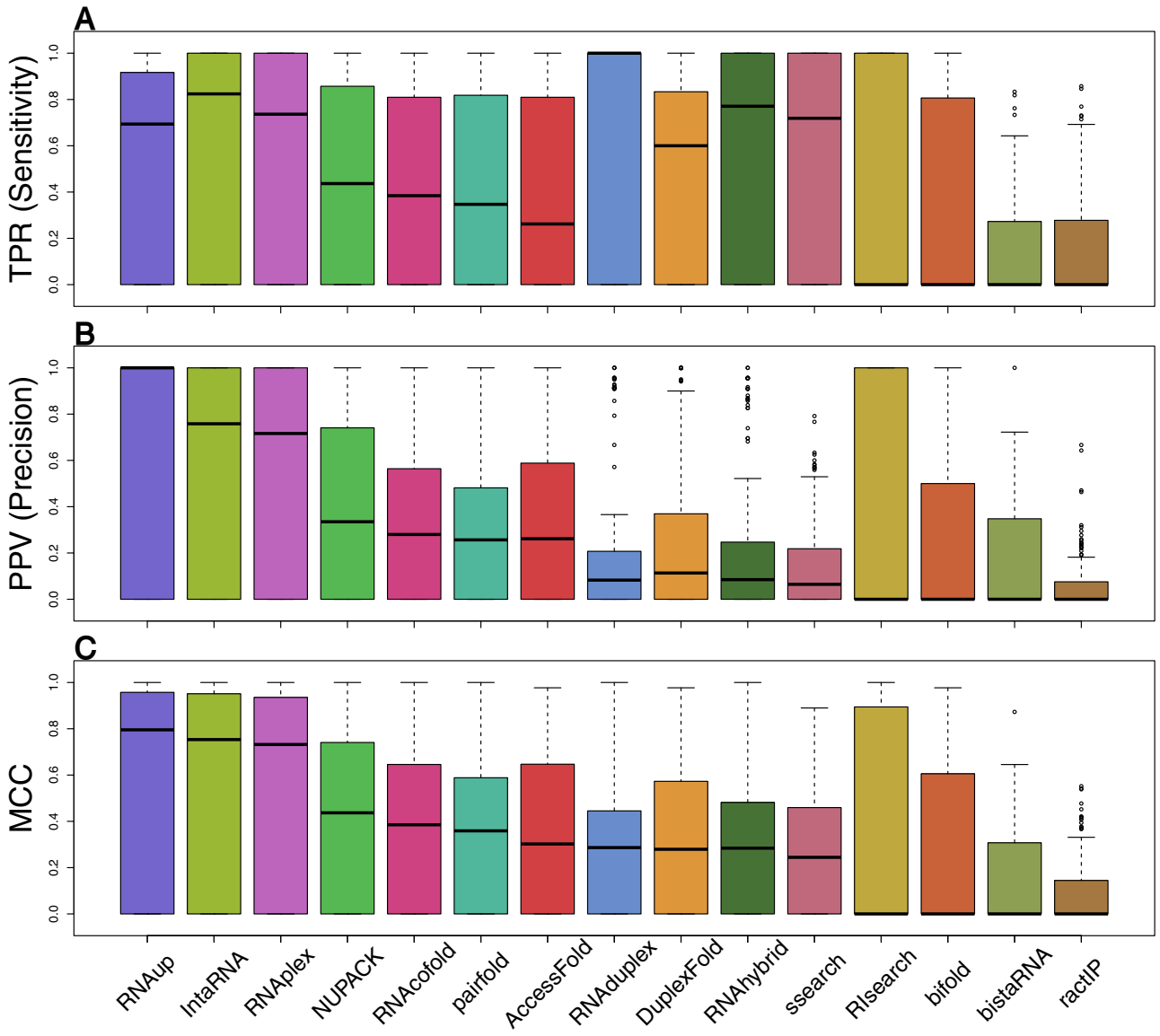

Fig. 1. The distribution of scores for RNA-RNA interaction prediction algorithms. (A) RNAduplex gave the highest median TPR (sensitivity) followed by IntaRNA. (B) RNAup was the most precise algorithm based on PPV score followed by the other accessibility based methods IntaRNA and RNAplex. (C) RNAup was the best prediction algorithm based on median MCC score, with IntaRNA and RNAplex giving similar scores. RactIP produced the worst overall MCC.

correct binding regions, functional characterization or identifying new targets (Chou *et al.*, 2015; Thomson *et al.*, 2011; Kuhn *et al.*, 2008; Vogel and Wagner, 2007). However, the incorrectly predicted human miRNAs (hsa-miR-21-5p, hsa-miR-29b-3p etc.) were validated by relatively strong evidence (Chou *et al.*, 2015), which could rule out this explanation.

RNA structure prediction (and also RNA-RNA interaction prediction) algorithms are based on biophysical assumptions where the influence of tertiary interactions and other factors are neglected (Mathews, 2006; Mathews and Turner, 2006; Wuchty *et al.*, 1999). RNA structures with the lowest free energy may not be the biologically active form, which may have multiple different conformations with different MFEs (Mathews, 2006; Mathews and Turner, 2006). Many algorithms ignore computationally expensive RNA structures (e.g. pseudoknots) (Hofacker *et al.*, 1994; Lorenz *et al.*, 2011; Do *et al.*, 2006). MFE methods also become inaccurate with longer RNA sequences (Mathews and Turner, 2006; Lange *et al.*, 2012; Lai and Meyer, 2015; Meyer, 2008). RNA interaction prediction algorithms generally do not consider multiple binding regions - only a few of which such as bistaRNA and ractIP, include multiple binding positions in their model (Poolsap *et al.*, 2011; Kato *et al.*, 2010). Cellular dynamics (i.e. interaction with other molecules, ion concentrations etc.) can influence RNA structures (Onoa and Tinoco, 2004) and RNA interactions (Mückstein *et al.*, 2006; Meyer, 2008), which is hard to factor into prediction models.

The *ssearch* tool uses the Smith-Waterman algorithm (Pearson and Lipman, 1988) and is the only pure alignment tool in our benchmark, although it is possible to use similar tools like BLAST or Blat to extract complementary regions for high-throughput predictions. Once the gap penalty and scoring matrix parameters were tweaked to make it more suitable for RNA-RNA interaction prediction, *ssearch* was quite successful and even comparable with some MFE methods (e.g. RNAhybrid and DuplexFold) (Figure 1).

Those MFE methods that include internal structures (e.g. pairfold, RNAcofold, bifold, NUPACK) are biased towards internal structures as many ncRNAs have stable internal structures (Clote et al., 2005). Therefore, using negative controls may lead to false significant predictions due to internal structures of interacting partners, giving misleading MFE

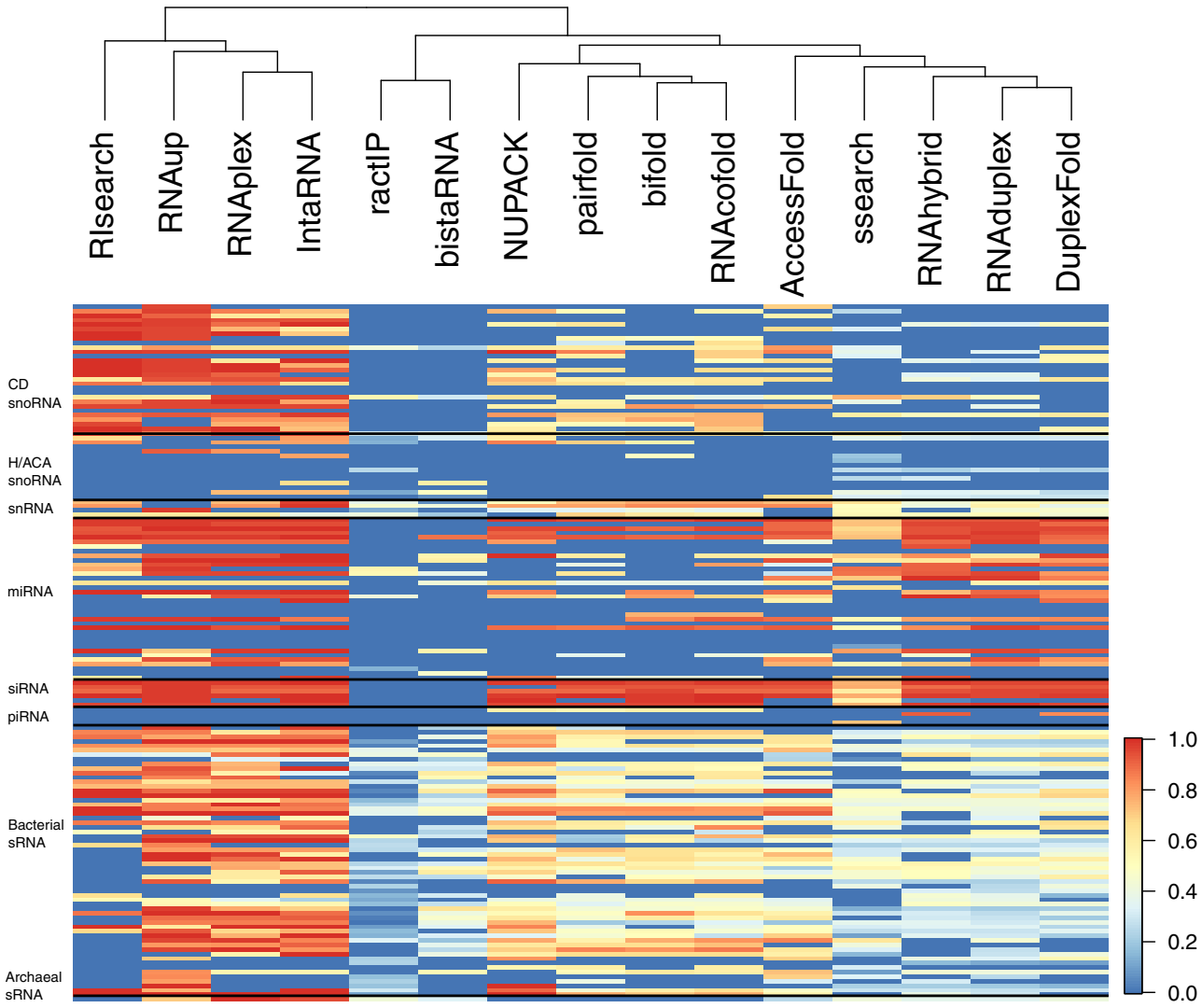

Fig. 2. This heatmap shows MCC values of each tool for entire dataset. The red cells display a higher MCC value denoting a better prediction. Similar methods are mostly clustered together based on these predictions (dendrogram at top). Row labels show the type of interactions. Predictions for the single archaeal sRNA are on the last row. An in depth examination of these results show that the algorithms are poor at predicting human miRNA-mRNA interactions (av. MCC: 0.22), snoRNAs (weaker for H/ACA as expected, av. MCC: 0.09), mouse piRNAs (av. MCC: 0.07). Conversely, they perform best on Arabidopsis miRNAs (av. MCC: 0.72), siRNAs (av. MCC: 0.71) and bacterial sRNAs (av. MCC: 0.40), which is most likely an effect of high complementarity in binding regions for these.

scores. We also observed this effect in our predictions (data not shown), and so excluded those algorithms from the significance test. They also have relatively slow running times, and some have problems utilizing memory (e.g. bifold). NUPACK is the best among this type of prediction methods and RNAcofold is the fastest (Table 1).

It is apparent that the algorithms do not necessarily perform equally for all types of RNA-RNA interactions, and it is better to select algorithms appropriate to the input dataset. For example, RIsearch is fast and accurate for eukaryotic datasets, and would be suitable for high throughput predictions which can be combined statistical significance testing of the predicted scores. IntaRNA and RNAplex seem to be reliable and relatively fast for all datasets. RNAup is precise and less prone to length bias (Supplementary Table S4).

4 Conclusion

Here we present one of the most comprehensive benchmark of RNA-RNA interaction prediction methods that covers almost all RNA-RNA interactions in RNA biology. We extended the previous work (Pain *et al.*, 2015; Lai and Meyer, 2015) by including all types of RNA-RNA interactions and the latest algorithms (DiChiacchio *et al.*, 2015) in the RNA interaction prediction field. We have included a test to determine the statistical significance of the predicted scores by each algorithm. We have also reported that increasing length of target RNAs which contain binding regions also negatively influences overall prediction quality (Supplementary Table S4).

Three accessibility based algorithms, RNAup, IntaRNA and RNAplex, performed best for all types of interactions. We found that the accessibility based MFE methods could also differentiate almost half of the native interactions from background in our bacterial dataset (Table 2). Therefore, carefully designed negative controls (e.g. dinucleotide shuffling) allow for the use of predicted MFE values and separate scores for native interactions from the background. This makes the accessibility algorithms ideal tools for *de novo* predictions, especially those with smaller run-times such as IntaRNA and RNAplex, since candidate target RNAs can be thousands of nts long. RNAplex is also effective on detecting correct interaction regions buried in larger RNA targets (Results and Supplementary Table S4).

RNA interaction prediction is still an expanding field. Advances in sequencing technology has unveiled a vast number of novel uncharacterized ncRNA transcripts in different clades of life. These methods are also showing that many ncRNAs utilize RNA-RNA interactions (Kudla *et al.*, 2011; Sharma *et al.*, 2016; Lu *et al.*, 2016) which makes RNA target prediction an important asset to determine functions of novel genes. Comparative methods are becoming popular (Wright *et al.*, 2013; Pain *et al.*, 2015; Lai and Meyer, 2015; Seemann *et al.*, 2011), and may increase the prediction accuracy (Wright *et al.*, 2013; Pain *et al.*, 2015). However, some other results suggest that there is little to be gained from comparative approaches for predicting interactions (Lai and Meyer, 2015; Richter and Backofen, 2012) due to low conservation of many ncRNAs (Lindgreen *et al.*, 2014). Unfortunately, most of the verified interactions in the RNA literature still belong to model species (human, *C. elegans, Arabidopsis* and *E. coli* etc.) which also raises the risk of overfitting results to a modest numbers of known interactions. Weak prediction rates for piRNAs may suggest inadequacy of prediction methods for novel regulatory RNAs, but even well-known miRNA interaction predictions have failed to be detected by any of the algorithms benchmarked (Figure 2). Archaeal regulatory systems are also not well studied, and only a handful of archaeal sRNAs have been identified. Therefore, non-comparative methods are still a robust way to produce *ab initio* interaction predictions. Our benchmark will help researchers to find an appropriate algorithm for functional annotation of unknown transcripts or a basis from which to improve or develop new methods. Our scripts and datasets are publicly available at Github (github.com/UCanCompBio/RNA_Interactions_Benchmark).

Acknowledgements and Funding

SUU is supported by a Biomolecular Interaction Centre and UC HPC (Bluefern) joint PhD Scholarship from the University of Canterbury. PPG is supported by Rutherford Discovery Fellowships, administered by the Royal Society of New Zealand. We also thank Lars Barquist for valuable discussions and comments.

Abbreviations

Non-coding RNA (ncRNA), microRNA (miRNA), small-interfering RNA (siRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), small Cajal body-specific RNA (scaRNA), piwi-interacting RNA (piRNA), messenger RNA (mRNA), small RNA (sRNA), long-noncoding RNAs (lncRNA), minimum free energy (MFE), true positive rate (TPR), positive predictive value (PPV), Matthews correlation coefficient (MCC), true positive (TP), false positive (FP), false negative (FN), true negative (TN), untranslated region (UTR), clustered regularly-interspaced short palindromic repeats (CRISPR) RNA (tracrRNA), ribonucleoprotein (RNP), average (av.), nucleotides (nts)

References

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the arabidopsis degradome. *Curr. Biol.*, **18**(10), 758–762.

Alkan, C., Karakoç, E., Nadeau, J. H., Sahinalp, S. C., and Zhang, K. (2006). RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**(2), 267–282.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–410.

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, **431**(7006), 350–355.

Ameres, S. L. and Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.*, **14**(8), 475–488.

Andronescu, M., Zhang, Z. C., and Condon, A. (2005). Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**(5), 987–1001.

Axtell, M. J., Westholm, J. O., and Lai, E. C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.*, **12**(4), 221.

Backofen, R. and Hess, W. R. (2010). Computational prediction of sRNAs and their targets in bacteria. *RNA Biol.*, **7**(1), 33–42.

Barquist, L. and Vogel, J. (2015). Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.*

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006). Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**(1), 3.

Bhaya, D., Davison, M., and Barrangou, R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273–297.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. *Cell*, **128**(6), 1089–1103.

Brown, J. W., Clark, G. P., Leader, D. J., Simpson, C. G., and Lowe, T. (2001). Multiple snoRNA gene clusters from arabidopsis. *RNA*, **7**(12), 1817–1832.

Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**(24), 2849–2856.

Cao, Y., Wu, J., Liu, Q., Zhao, Y., Ying, X., Cha, L., Wang, L., and Li, W. (2010). sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA*, **16**(11), 2051–2057.

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**(4), 642–655.

Chen, W.-H., van Noort, V., Lluch-Senar, M., Hennrich, M. L., H Wodke, J. A., Yus, E., Alibés, A., Roma, G., Mende, D. R., Pesavento, C., Typas, A., Gavin, A.-C., Serrano, L., and Bork, P. (2016). Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic Acids Res.*, **44**(3), 1192–1202.

Chen, X. (2008). MicroRNA metabolism in plants. *Curr. Top. Microbiol. Immunol.*, **320**, 117–136.

Chitsaz, H., Backofen, R., and Cenk Sahinalp, S. (2009a). biRNA: Fast RNA-RNA binding sites prediction. In S. L. Salzberg and T. Warnow, editors, *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 25–36. Springer Berlin Heidelberg.

Chitsaz, H., Salari, R., Sahinalp, S. C., and Backofen, R. (2009b). A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, **25**(12), i365–73.

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., Tsai, T.-R., Ho, S.-Y., Jian, T.-Y., Wu, H.-Y., Chen, P.-R., Lin, N.-C., Huang, H.-T., Yang, T.-L., Pai, C.-Y., Tai, C.-S., Chen, W.-L., Huang, C.-Y., Liu, C.-C., Weng, S.-L., Liao, K.-W., Hsu, W.-L., and Huang, H.-D. (2015). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*

Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., and Sorek, R. (2016). Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res.*

Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *Plant Cell*, **23**(2), 431–442.

Darzacq, X., Jády, B. E., Verheggen, C., Kiss, A. M., Bertrand, E., and Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2'-o-methylation and pseudouridylation guide RNAs. *EMBO J.*, **21**(11), 2746–2756.

Deleavey, G. F. and Damha, M. J. (2012). Designing chemically modified oligonucleotides for targeted gene silencing. *Chem. Biol.*, **19**(8), 937–954.

DiChiacchio, L., Sloma, M. F., and Mathews, D. H. (2015). AccessFold: Predicting RNA-RNA interactions with consideration for competing Self-Structure. *Bioinformatics*.

Dieterich, C. and Stadler, P. F. (2012). Computational biology of RNA interactions. *Wiley Interdiscip. Rev. RNA*.

Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and Pierce, N. A. (2007). Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**(1), 65–88.

Do, C. B., Woods, D. A., and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**(14), e90–8.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**(10), e1002195.

Gardner, P. P., Bateman, A., and Poole, A. M. (2010). SnoPatrol: how many snoRNA genes are there? *J. Biol.*, **9**(1), 4.

Gerlach, W. and Giegerich, R. (2006). GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics*, **22**(6), 762–764.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, **319**(6055).

Gorodkin, J., Stricklin, S. L., and Stormo, G. D. (2001). Discovering common stem–loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**(10), 2135–2144.

Gottesman, S. (2004). The small RNA regulators of escherichia coli: roles and mechanisms*. *Annu. Rev. Microbiol.*, **58**, 303–328.

Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.-P., Zhou, Y., Kang, J.-Y., Wang, X., Li, H., Hua, M.-M., Zhao, S., Hu, S.-D., Wu, L.-G., Shi, H.-J., Li, Y., Fu, X.-D., Qu, L.-H., Wang, E.-D., and Liu, M.-F. (2015). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res.*, **25**(2), 266.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., and Serrano, L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**(5957), 1268–1271.

Güell, M., Yus, E., Lluch-Senar, M., and Serrano, L. (2011). Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.*, **9**(9), 658–669.

Gumbel, E. J. (1958). Statistics of extremes. 1958. *Columbia Univ. press, New York*.

Hodas, N. O. and Aalberts, D. P. (2004). Efficient computation of optimal oligo-RNA binding. *Nucleic Acids Res.*, **32**(22), 6636–6642.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte fi¿½r Chemie Chemical Monthly*, **125**(2), 167–188.

Holmqvist, E. and Vogel, J. (2013). A small RNA serving both the hfq and CsrA regulons. *Genes Dev.*, **27**(10), 1073–1078.

Jäger, D., Pernitzsch, S. R., Richter, A. S., Backofen, R., Sharma, C. M., and Schmitz, R. A. (2012). An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains. *Nucleic Acids Res.*, **40**(21), 10964–10979.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biol.*, **2**(11), e363.

Karijolich, J. and Yu, Y.-T. (2010). Spliceosomal snRNA modifications and their function. *RNA Biol.*, **7**(2), 192–204.

Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K., and Akutsu, T. (2010). RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**(18), i460–6.

Kehr, S., Bartschat, S., Stadler, P. F., and Tafer, H. (2011). PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics*, **27**(2), 279–280.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.*

Kery, M. B., Feldman, M., Livny, J., and Tjaden, B. (2014). TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Res.*, **42**(Web Server issue), W124–9.

Kidner, C. A. and Martienssen, R. A. (2005). The developmental role of microRNA in plants. *Curr. Opin. Plant Biol.*, **8**(1), 38–44.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**(10), 1165–1178.

Kiss, A. M., Jady, B. E., Bertrand, E., and Kiss, T. (2004). Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol.*, **24**(13), 5797–5807.

Kiss, T. (2002). Small nucleolar RNAs. *Cell*, **109**(2), 145–148.

Klattenhoff, C. and Theurkauf, W. (2008). Biogenesis and germline functions of piRNAs. *Development*, **135**(1), 3–9.

Kozomara, A. and Griffiths-Jones, S. (2013). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, pages gkt1181–.

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proceedings of the National Academy of Sciences*, **108**(24), 10010–10015.

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, Jr, A. V., Nuovo, G. J., and Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods*, **44**(1), 47–54.

Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics*, **193**(3), 651–669.

Lai, D. and Meyer, I. M. (2015). A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic Acids Res.*

Lange, S. J., Maticzka, D., Möhl, M., Gagnon, J. N., Brown, C. M., and Backofen, R. (2012). Global or local? predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**(12), 5215–5226.

Lestrade, L. and Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**(Database issue), D158–62.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**(1), 15–20.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N. E., Biggs, P. J., Thomson, N. R., Barquist, L., Poole, A. M., and Gardner, P. P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput. Biol.*, **10**(10), e1003907.

Lorenz, R., Bernhart, S. H., Siederdissen, C. H. Z., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., Mesirov, J. P., Cech, T. R., and Chang, H. Y. (2016). RNA duplex map in living cells reveals Higher-Order transcriptome structure. *Cell*.

Markham, N. R. and Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.

Mathews, D. H. (2006). Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**(3), 526–532.

Mathews, D. H. and Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**(3), 270–278.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**(2), 442–451.

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat. Rev. Genet.*, **5**(4), 316–323.

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genet.*, **5**(4), e1000459.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**(6-7), 1105–1119.

Meyer, I. M. (2008). Predicting novel RNA–RNA interactions. *Curr. Opin. Struct. Biol.*, **18**(3), 387–393.

Millar, A. A. and Waterhouse, P. M. (2005). Plant and animal microRNAs: similarities and differences. *Funct. Integr. Genomics*, **5**(3), 129–135.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. (2006). Thermodynamics of RNA–RNA binding. *Bioinformatics*, **22**(10), 1177–1182.

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. U. S. A.*, **77**(11), 6309–6313.

Omer, A. D., Lowe, T. M., Russell, A. G., Ebhardt, H., Eddy, S. R., and Dennis, P. P. (2000). Homologs of small nucleolar RNAs in archaea. *Science*, **288**(5465), 517–522.

O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-Seq capacity. *Curr. Protoc. Mol. Biol.*, pages 4–19.

Onoa, B. and Tinoco, Jr, I. (2004). RNA folding and unfolding. *Curr. Opin. Struct. Biol.*, **14**(3), 374–379.

Oğul, H., Umu, S. U., Tuncel, Y. Y., and Akkaya, M. S. (2011). A probabilistic approach to microRNA-target binding. *Biochem. Biophys. Res. Commun.*, **413**(1), 111–115.

Pain, A., Ott, A., Amine, H., Rochat, T., Bouloc, P., and Gautheret, D. (2015). An assessment of bacterial small RNA target prediction programs. *RNA Biol.*, **12**(5), 509–513.

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**(3), 635–650.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.*, **85**(8), 2444–2448.

Peer, A. and Margalit, H. (2011). Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions. *J. Bacteriol.*, **193**(7), 1690–1701.

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. *Science*, **337**(6099), 1159, 1161.

Pervouchine, D. D. (2004). IRIS: intermolecular RNA interaction search. *Genome Inform.*, **15**(2), 92–101.

Piekna-Przybylska, D., Decatur, W. A., and Fournier, M. J. (2007). New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA*, **13**(3), 305–312.

Poolsap, U., Kato, Y., Sato, K., and Akutsu, T. (2011). Using binding profiles to predict binding sites of target RNAs. *J. Bioinform. Comput. Biol.*, **9**(6), 697–713.

Prasse, D., Ehlers, C., Backofen, R., and Schmitz, R. A. (2013). Regulatory RNAs in archaea: first target identification in methanoarchaea. *Biochem. Soc. Trans.*, **41**(1), 344–349.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**(10), 1507–1517.

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. *Nat. Biotechnol.*, **22**(3), 326–330.

Richter, A. S. and Backofen, R. (2012). Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA Biol.*, **9**(7), 954–965.

Seemann, S. E., Richter, A. S., Gesell, T., Backofen, R., and Gorodkin, J. (2011). PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, **27**(2), 211–219.

Sharma, C. M. and Vogel, J. (2014). Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol.*, **19**, 97–105.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010). The primary transcriptome of the major human pathogen helicobacter

pylori. *Nature*, **464**(7286), 250–255.

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B. J. (2016). Global mapping of human RNA-RNA interactions. *Mol. Cell*, **62**(4), 618–626.

Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, **43**(6), 880–891.

Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**(22), 2657–2663.

Tafer, H., Kehr, S., Hertel, J., Hofacker, I. L., and Stadler, P. F. (2010). RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics*, **26**(5), 610–616.

Thébault, P., Bourqui, R., Benchimol, W., Gaspin, C., Sirand-Pugnet, P., Uricaru, R., and Dutour, I. (2014). Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. *Brief. Bioinform.*

Thomson, D. W., Bracken, C. P., and Goodall, G. J. (2011). Experimental strategies for microRNA target identification. *Nucleic Acids Res.*, **39**(16), 6845–6853.

Tjaden, B. (2008). TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res.*, **36**(Web Server issue), W109–13.

Vogel, J. (2009). A rough guide to the non-coding RNA world of salmonella. *Mol. Microbiol.*, **71**(1), 1–11.

Vogel, J. and Luisi, B. F. (2011). Hfq and its constellation of RNA. *Nat. Rev. Microbiol.*, **9**(8), 578–589.

Vogel, J. and Wagner, E. G. H. (2007). Target identification of small noncoding RNAs in bacteria. *Curr. Opin. Microbiol.*, **10**(3), 262–270.

Waters, L. S. and Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*, **136**(4), 615–628.

Wenzel, A., Akbaşli, E., and Gorodkin, J. (2012). RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, **28**(21), 2738–2746.

Will, C. L. and Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**(7).

Witkos, T. M., Koscianska, E., and Krzyzosiak, W. J. (2011). Practical aspects of microRNA target prediction. *Curr. Mol. Med.*, **11**(2), 93–109.

Workman, C. and Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**(24), 4816–4822.

Wright, P. R., Richter, A. S., Papenfort, K., Mann, M., Vogel, J., Hess, W. R., Backofen, R., and Georg, J. (2013). Comparative genomics boosts target prediction for bacterial small RNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **110**(37), E3487–96.

Wuchty, S., Fontana, W., Hofacker, I. L., and Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**(2), 145–165.

Yang, Y., Wang, Y.-P., and Li, K.-B. (2008). MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics*, **9 Suppl 12**, S4.

Yoshihama, M., Nakao, A., and Kenmochi, N. (2013). snOPY: a small nucleolar RNA orthological gene database. *BMC Res. Notes*, **6**, 426.

Zhang, P., Kang, J.-Y., Gou, L.-T., Wang, J., Xue, Y., Skogerboe, G., Dai, P., Huang, D.-W., Chen, R., Fu, X.-D., Liu, M.-F., and He, S. (2015). MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res.*, **25**(2), 193–207.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M. Q., and Chen, R. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**(D1), D203–8.

Zuker, M. (2000). Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**(3), 303–310.

Zuker, M. and Sankoff, D. (1984). RNA secondary structures and their prediction. *Bltn Mathcal Biology*, **46**(4), 591–621.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**(1), 133–148.

CHAPTER IV - The RNA Avoidance Hypothesis

Transcription is the biosynthesis of RNA molecules using the data encoded in a DNA molecule. It is one of the fundamental processes of all living cells with translation, which is the biosynthesis of protein using the information transcribed in mRNAs. DNA-dependent holoenzyme RNA polymerase starts transcription initiation by binding to specific promoter regions (Browning & Busby 2004; Lee et al. 2012; Reznikoff et al. 1985) in bacteria. This is followed by unwinding of the DNA duplex and formation of transcripts (Lee et al. 2012). The translation initiation involves assembly of initiation factors and ribosomal SSUs on an mRNA, which is followed by detection of the start codon by the initiator tRNA (Kozak 1999). The elongation phase involves the movement of the ribosome along the mRNA strand (Schmeing & Ramakrishnan 2009). The translation termination stage involves the release of polypeptide chain and disassembly of ribosomal subunits (Schmeing & Ramakrishnan 2009). In prokaryotes, transcription and translation of mRNAs occur in same place without any separation by internal membranes (Gowrishankar & Harinarayanan 2004), and even nascent mRNAs are translated during transcription (Laursen et al. 2005). However, protein and mRNA levels are often poorly correlated in both prokaryotes and eukaryotes (de Sousa Abreu et al. 2009; Vogel & Marcotte 2012; Kwon et al. 2014; Maier et al. 2011; Lu et al. 2007; Taniguchi et al. 2010; Chen et al. 2016), which is a major barrier to precision bioengineering and quantification of protein levels. Furthermore, synonymous mutations do not alter the translated protein sequence, but can significantly influence the protein abundances (Tuller et al. 2010; Parmley & Hurst 2007). There are two well-known global factors that influence protein expression (Tuller et al. 2010), which explain some variation in mRNA and protein levels. These two factors are codon usage (Tuller et al. 2010; Ikemura 1981; Ikemura 1985; Akashi 1994) and secondary structure of mRNAs (Chamary & Hurst 2005; Tuller et al. 2010; Gaspar et al. 2013; Pelletier & Sonenberg 1987; Gu et al. 2014). On the other hand, these features account for only half of variation, and sometimes they can explain even less (Kudla et al. 2009; Maier et al. 2011; Plotkin & Kudla 2011;

Goodman et al. 2013; Chen et al. 2016).

The aim of the following study is to examine the unexplained variance between mRNA and protein expression levels. In order to achieve this goal, we carefully examined the crosstalk interactions among core ncRNAs and mRNAs of prokaryotic cells. Despite a growing recognition of the importance of RNA-RNA mediated regulation in prokaryotes (Waters & Storz 2009; Storz et al. 2011; Updegrove et al. 2015), no study has yet evaluated the significance of unfavorable interactions between RNAs.

We have proposed that crosstalk is selectively disadvantageous, therefore this may be detectable as 'an avoidance signal'. In other words, RNA-RNA interactions leading to crosstalk should significantly influence protein output and be underrepresented in prokaryotic genomes. We have tested this *in silico* on available bacterial and archaeal genomes by extracting core mRNAs and core ncRNA genes. We also collected available prokaryotic gene expression data to detect crosstalk influence on protein expression. Furthermore, we designed 13 green fluorescence protein (GFP) reporter mRNAs and inserted these into *E. coli*.

Our results show that crosstalk avoidance is a widespread phenomenon in bacteria and archaea, and is supported by an evolutionarily conserved signature. Our GFP reporter assay results show that we can accurately control gene expression levels by accounting for mRNA-ncRNA avoidance. These results imply that RNA avoidance is an important factor for the gene optimization problem and prokaryotic translation models.

The following manuscript describing these results in detail is in review at the journal *Elife*. A latest version of the preprint can be accessed from this link : http://dx.doi.org/10.1101/033613 .

REFERENCES

Akashi, H., 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3), pp.927–935.

Browning, D.F. & Busby, S.J., 2004. The regulation of bacterial transcription initiation. *Nature reviews. Microbiology*, 2(1), pp.57–65.

Chamary, J.V. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9), p.R75.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Gaspar, P. et al., 2013. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic acids research*, 41(6), pp.e73–e73.

Goodman, D.B., Church, G.M. & Kosuri, S., 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(6157), pp.475–479.

Gowrishankar, J. & Harinarayanan, R., 2004. Why is transcription coupled to translation in bacteria? *Molecular microbiology*, 54(3), pp.598–603.

Gu, W. et al., 2014. The impact of RNA structure on coding sequence evolution in both bacteria and eukaryotes. *BMC evolutionary biology*, 14, p.87.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), pp.13–34.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology*, 151(3), pp.389–409.

Kozak, M., 1999. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234(2), pp.187–208.

Kudla, G. et al., 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* , 324(5924), pp.255–258.

Kwon, T. et al., 2014. Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. *Journal of proteome research*, 13(5), pp.2370–2380.

Laursen, B.S. et al., 2005. Initiation of protein synthesis in bacteria. *Microbiology and molecular biology reviews: MMBR*, 69(1), pp.101–123.

Lee, D.J., Minchin, S.D. & Busby, S.J.W., 2012. Activating transcription in bacteria. *Annual review of microbiology*, 66, pp.125–152.

Lu, P. et al., 2007. Absolute protein expression profiling estimates the relative contributions of

transcriptional and translational regulation. *Nature biotechnology*, 25(1), pp.117–124.

Maier, T. et al., 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular systems biology*, 7, p.511.

Parmley, J.L. & Hurst, L.D., 2007. How do synonymous mutations affect fitness? *BioEssays: news and reviews in molecular, cellular and developmental biology*, 29(6), pp.515–519.

Pelletier, J. & Sonenberg, N., 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 65(6), pp.576–581.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1), pp.32–42.

Reznikoff, W.S. et al., 1985. The regulation of transcription initiation in bacteria. *Annual review of genetics*, 19, pp.355–387.

Schmeing, T.M. & Ramakrishnan, V., 2009. What recent ribosome structures have revealed about the mechanism of translation. *Nature*, 461(7268), pp.1234–1242.

de Sousa Abreu, R. et al., 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12), pp.1512–1526.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.

Tuller, T. et al., 2010. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8), pp.3645–3650.

Updegrove, T.B., Shabalina, S.A. & Storz, G., 2015. How do base-pairing small RNAs evolve? *FEMS microbiology reviews*, 39(3), pp.379–391.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea

Authors:

Sinan U. Umu[1,2], Anthony M. Poole[1,2,3], Renwick C.J. Dobson[1,2,5], Paul P. Gardner[1,2,4]*

[1]School of Biological Sciences, [2]Biomolecular Interaction Centre, [3]Allan Wilson Centre for Molecular Ecology & Evolution, [4]Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand

[5]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, VIC 3010, Australia

*paul.gardner@canterbury.ac.nz

Abstract

A critical assumption of gene expression analysis is that mRNA abundances broadly correlate with protein abundance, but these two are often imperfectly correlated. Some of the discrepancy can be accounted for by two important mRNA features: codon usage and mRNA secondary structure. We present a new global factor, called mRNA:ncRNA avoidance, and provide evidence that avoidance increases translational efficiency. We also demonstrate a strong selection for avoidance of stochastic mRNA:ncRNA interactions across prokaryotes, and that these have a greater impact on protein abundance than mRNA structure or codon usage. By generating synonymously variant green fluorescent protein (GFP) mRNAs with different potential for mRNA:ncRNA interactions, we demonstrate that GFP levels correlate well with interaction avoidance. Therefore, taking stochastic mRNA:ncRNA interactions into account enables precise modulation of protein abundance.

Main Text

Introduction

It should in principle be possible to predict protein abundance from genomic data. However,

protein and mRNA levels are not strongly correlated (de Sousa Abreu et al. 2009; Vogel & Marcotte 2012; Kwon et al. 2014; Maier et al. 2011; Lu et al. 2007; Taniguchi et al. 2010; Chen et al. 2016), which is a major barrier to precision bioengineering and quantification of protein levels. mRNA secondary structure (Pelletier & Sonenberg 1987; Chamary & Hurst 2005), codon usage (Ikemura 1981; Sharp & Li 1987; Andersson & Kurland 1990), and mRNA (and protein) degradation rates (Maier et al. 2011) are commonly invoked to explain this discrepancy (Boël et al. 2016). Yet, at best, these features account for only 40% of variation, and in some instances explain very little of the observed variation (Kudla et al. 2009; Maier et al. 2011; Plotkin & Kudla 2011; Goodman et al. 2013; Chen et al. 2016). Here we show that crosstalk interactions between ncRNAs and mRNAs also impact protein abundance, and that such interactions have a greater effect than either mRNA secondary structure or codon usage. We measured interactions between a set of evolutionarily conserved core mRNAs and ncRNAs from 1,700 prokaryotic genomes using minimum free energy (MFE) models. For 97% of species, we find a reduced capacity for interaction between native RNAs relative to controls. Furthermore, by generating synonymously variant GFP mRNAs that differ in their potential to interact with core ncRNAs, we demonstrate that GFP expression levels can be both predicted and controlled. Our results demonstrate that there is strong selection for avoidance of stochastic mRNA:ncRNA interactions across prokaryotes. Applying this knowledge to mRNA design will enable precise control of protein abundance through the incorporation or exclusion of inhibitory interactions with native ncRNAs.

Result and Discussion

To examine if avoidance of stochastic mRNA:ncRNA interactions is a feature of transcriptomes in bacteria and archaea, we estimated the strength of all possible intermolecular RNA interactions using a minimum free energy (MFE) model (Mückstein et al. 2006) using core ncRNAs and mRNAs. In this work the core ncRNAs are six well conserved and highly expressed tRNA, rRNA, RNase P RNA, SRP RNA, tmRNA and 6S RNA families annotated by Rfam (Gardner et al. 2011; Nawrocki et al. 2015), the core mRNAs are 114 well conserved mRNAs found across bacteria, 40 of which are also conserved across archaea (Wu et al. 2013).

If stochastic interactions are selected against, because of the capacity for abundant ncRNAs (Lindgreen et al. 2014; Deutscher 2006; Giannoukos et al. 2012) to impact translation (Waters & Storz 2009; Storz et al. 2011), such negative selection would be most comparable between species and readily detected for broadly conserved ncRNAs and mRNAs. Under-representation of interactions has been considered for the specific case of Shine-Dalgarno-like (SD-like) sequences and the ribosome (Li et al. 2012; Woolstenhulme et al. 2015; Borg & Ehrenberg 2015; Diwan & Agashe 2016) and between microRNAs and 3` UTRs (Bartel & Chen 2004; Farh et al. 2005; Stark et al. 2005; van Dongen et al. 2008). We computed the free energy distribution of interactions between highly conserved mRNA:ncRNA pairs and compared this to a number of negative control interactions, which serve to show the expected distribution of binding energy values (Figure 1A). The initiation of translation has been shown to be the rate limiting step for translation (Tuller & Zur 2015; Plotkin & Kudla 2011; Nakahigashi et al. 2014), therefore, we focus our analysis on the first 21 nucleotides of the mRNA coding sequence (CDS). This has the further advantage of reducing computational complexity. We also test a variety of negative control mRNA regions, which are unlikely to play a functional role in RNA:ncRNA interactions. The mRNA controls include (1) di-nucleotide preserving shuffled sequences (Workman & Krogh 1999) (orange, Figure 1A), (2) homologous mRNAs from another phylum (with a compatible guanine-cytosine (G+C) content) (purple), (3) downstream regions 100 base pairs (bps) within the CDS (pink), (4) the reverse complement of the 5` of CDSs (green), and lastly (5) unannotated (intergenic) genomic regions (yellow). Our interaction predictions in a single model strain show that native interactions consistently have higher (i.e. less stable) free energies than expected when compared to the five different mRNA negative controls: that is, there is a reduced capacity for native mRNAs and native ncRNAs to interact. We also compared different energy models and confirm that the MFE shift is a result of intermolecular binding (Figure 1-figure supplement 1A,B,C). We subsequently deployed the most conservative negative control (i.e. di-nucleotide preserving shuffle) and free energy model (Figure 1-figure supplement 1C) to detect if this shift for less stable binding of mRNA:ncRNA is true of all bacteria and archaea.

In terms of stoichiometry, the model we use assumes that ncRNA expression levels are vastly in excess of mRNA expression levels (i.e. [ncRNA]>>[mRNA]) (Giannoukos et al. 2012; Deutscher 2006). This is generally a biologically reasonable assumption when focussing on core genes based upon past analysis and our own work with RNA-seq data from a range of bacteria and archaea (Figure 4) (Lindgreen et al. 2014). Consequently, any potential mRNA interaction regions are saturated with ncRNA, therefore a summative model of interaction energies is a reasonable approximation to the estimated impact of excess hybridization. If modelling ncRNAs that are not so abundant, then a model weighted by expression level may be advantageous, but it is difficult to assess these across all conditions and developmental stages that are evolutionarily relevant. In order to ensure that our analysis is comparable across all bacteria and archaea we have focussed on just the most highly conserved ncRNA and protein-coding genes. Although, many of the ncRNAs are highly structures and are bound by RNA-binding proteins this is not the case during either synthesis and degradation of these products, furthermore, a fraction of the RNA components of these genes will be exposed. Therefore we expect these will form useful datasets for initial testing of our hypothesis.

In order to assess whether mRNA:ncRNA is an evolutionarily conserved phenomenon, we calculated intermolecular binding energies for conserved ncRNAs and mRNAs from 1,582 bacterial and 118 archaeal genomes and compared these to a negative control dataset derived using a di-nucleotide frequency preserving shuffling procedure (Workman & Krogh 1999). This measures a property that we call the `extrinsic avoidance` of mRNA:ncRNA interactions, yet this approach may fail to identify genuine avoidance in cases when the G+C content differences between interacting RNAs is extreme. Measuring only extrinsic avoidance (using shuffled mRNAs as negative controls), we found that stochastic mRNA:ncRNA interactions are significantly underrepresented in most (73%) of the prokaryotic phyla ($P < 0.05$, one-tailed Mann-Whitney U test) (Figure 1B,C and Figure 1-figure supplement 2). This indicates that there is selection against stochastic interactions in both bacteria and archaea.

We next sought to establish the degree to which intrinsic G+C features of RNAs lead to

avoidance of stochastic interactions (Figure 1D). A similar idea has been proposed which suggests that purine loading in thermophilic bacteria may limit mRNA:mRNA interactions (Lao & Forsdyke 2000). A test of G+C composition revealed a significant difference ($P < 0.05$, two-tailed Mann-Whitney U test) between mRNAs and ncRNAs for 95% of bacteria and archaea (Figure 1D,E). Therefore, either extrinsic or intrinsic avoidance signals indicate that selection against stochastic interactions and it is near-universal for the prokaryotes (97% of all strains) (Figure 1E and Supplementary file 1A,B).

Our results clearly establish a signature of selection that acts to minimise stochastic mRNA:ncRNA interactions. However, with thousands of potential interacting RNA species in even simple prokaryotic systems (Vivancos et al. 2010; Sharma et al. 2010), the complete avoidance of stochastic interactions is combinatorially unlikely. Therefore there ought to be a tradeoff between avoidance and optimal expression. To assess this, we examined the relationship between potential stochastic interactions and the variation between mRNA and cognate protein levels for four previously published endogenous mass spectrometry datasets from *Escherichia coli (E. coli)* and *Pseudomonas aeruginosa (P. aeruginosa)* (Laurent et al. 2010; Kwon et al. 2014; Lu et al. 2007). We computed Spearman's correlation coefficients between protein abundances and extrinsic avoidance, 5` end internal mRNA secondary structure and codon usage. Of the three measures, avoidance is significantly correlated in all four datasets (Spearman's rho values are between 0.11 - 0.17 and corresponding *P* values are between 0.01 - $1.3 \times 10^{-12}$). In contrast, 5` end mRNA structure significantly correlates in two datasets, and codon usage significantly correlates in all four datasets. This indicates that, despite strong selection against stochastic interactions, such interactions do significantly impact the proteome (Figure 2A and Supplementary file 3). We have also conducted an "outlier analysis" on one of the *E. coli* datasets (Laurent et al. 2010). We have selected the top and bottom-most expressed genes relative to mRNA expression levels and computed Z-scores for each of codon-usage, internal secondary structure and avoidance measures. We found that avoidance measures shows the most extreme shifts downwards for the bottom-most expressed genes and is shifted the highest for the top-most genes (Figure 2-figure supplement 4).

We also test how mRNA:ncRNA crosstalk impacts the translation of transformed mRNAs that have not coevolved with the ncRNA repertoire (low avoidance mRNAs are rare in native datasets). We examined two available *E. coli*-based GFP experimental datasets (Goodman et al. 2013; Kudla et al. 2009), where synonymous mRNAs are generated for a GFP reporter gene. This enables the assessment of the impact of synonymous changes on protein abundance using fluorescence. Avoidance and mRNA secondary structure are both significantly correlated with fluorescence, whereas codon usage is not (Spearman's rho values are 0.11 and 0.65 , the corresponding *P* values are $3.17x10^{-41}$ and 1.69^{-20}) (Figure 2A). Note that one of the GFP datasets (Goodman et al. 2013) uses native *E. coli* mRNA 5` ends for their constructs, whereas the other GFP dataset (Kudla et al. 2009) is randomly generated. We observe that the influence of avoidance on gene expression for randomly sampled synonymous mRNAs is strong (Figure 2-figure supplement 3), while endogeneous gene expression is limited. Presumably, due to negative selection pruning low avoidance mRNAs from the gene pool (Figure 2A).

For each of the seven datasets described above we have tested linear models of measures of mRNA levels, codon usage, internal secondary structure and avoidance (Figure 2-figure supplement 3 and Supplementary file 5). Avoidance alone explains around 35% of variance in GFP datasets where extreme mRNA compositions can be explored, whereas in native mass-spec derived datasets 2-3% of the variance is explained by avoidance alone. Codon usage describes 2% to -0.5% of variance in GFP data, and 19% to 0.3% of variance in mass-spec derived datasets. Internal secondary structure 33% to 10% in GFP datasets, and 0.2% to 0% of the variance in mass-spec derived datasets. Using all four measures in combination across the seven datasets between 70% and 42% of variation in protein levels can be explained, removing avoidance from the model reduces these estimates by between 56% and 0.7%. Thus, avoidance is at least as good an explanation of variation in protein abundance as either codon usage and internal mRNA secondary structure.

Our results indicate that crosstalk between mRNAs and ncRNAs can impact protein expression

levels. We therefore predict that taking crosstalk into account will enable the design of constructs where protein expression levels can be precisely controlled. To test this, we generated GFP constructs based on the following constraints: codon bias, 5` end mRNA secondary structure stability and crosstalk avoidance (see 'Materials and methods'). Our constructs are designed to capture the extremes of one variable, while controlling other variables (e.g. high or low avoidance and near-average codon bias and mRNA secondary structure). The G+C content, a known confounding factor, was also strictly controlled for each construct. We selected a commercial service to perform our GFP transformations to avoid possible bias and increased the robustness of our approach (Ioannidis & Khoury 2011). We predicted that a construct where all three parameters are optimised will result in a high expression. Consistent with predictions, our optimised construct had maximal expression (Figure 2-figure supplement 1). Of the three parameters, avoidance showed the largest range, suggesting that tuning this parameter permits expression levels to be finely controlled ($R_s = 0.56$, $P = 6.9 \times 10^{-6}$) (Figure 2B,C,D and Figure 2-figure supplement 1-4).

For a final confirmation of the avoidance hypothesis, we tested the *Thermus thermophilus (T. thermophilus)* HB8 SSU ribosomal RNA, which is a component of one of the most complete prokaryotic ribosomal structures available in the PDB (Rozov et al. 2015). We identified the regions of the SSU rRNA that had the least capacity to interact with *T. thermophilus* core mRNAs and found that these regions were generally not bound to either ribosomal proteins or other ncRNAs, such as the LSU rRNA ($P = 2.49 \times 10^{-17}$, Fisher's exact test) (Figure 3; see 'Materials and methods'). The influence of internal SD-like regions on translation pausing have been described elsewhere (Li et al. 2012), in addition we note that the anti-SD region on SSU rRNA is one of the RNA avoidance regions (Figure 3A).

This study focusses on the 5` ends of the CDS as this region is important for the initiation of translation (Plotkin & Kudla 2011; Tuller & Zur 2015) and is a consistent feature of all the genomic, transcriptomic, proteomic and GFP expression datasets that we have evaluated in this work. In smaller-scale tests we have observed similar conserved avoidance signals within the

entire CDSs (Figure 3-figure supplement 1) and within the 5`UTRs (Figure 3-figure supplement 2). Furthermore, we predict that similar signals can be observed for mRNA:mRNA and ncRNA:ncRNA avoidance. Although the impacts of these features are challenging to validate, interactions between clustered regularly interspaced short palindromic repeats (CRISPR) spacer sequences (Bhaya et al. 2011) and core ncRNAs are good candidates to test ncRNA:ncRNA avoidance.

In conclusion, our results indicate that the specificity of prokaryotic ncRNAs for target mRNAs is the result of selection both for a functional interaction and against stochastic interactions. Our experimental results support the view that stochastic interactions are selected against, due to deleterious outcomes on expression. We suspect avoidance of crosstalk interactions has several evolutionary consequences. First, as transcriptional outputs become more diverse in evolution, we expect that the probability of stochastic interactions for both new ncRNAs and mRNAs becomes higher. This will impact the emergence of new, high abundance RNAs, since selection for high abundance may be mitigated by deleterious crosstalk events. Second, we predict that stochastic interactions limit the number of simultaneously transcribed RNAs, since the combinatorics of RNA:RNA interactions imply that eventually stochastic interactions cannot be avoided. This may in turn drive selection for forms of spatial or temporal segregation of transcripts. Finally, taking codon usage, mRNA secondary structure and potential mRNA:ncRNA interactions into account allows better prediction of proteome outputs from genomic data, and informs the precise control of protein levels via manipulation of synonymous mRNA sequences (Figure 2-figure supplement 5).

Materials and Methods

Here we summarize the data sources, materials and methods corresponding to our manuscript. We performed all statistical analyses in R, and all other computational methods in Python 2.7 or Bash shell scripts. We explicitly cite all the bioinformatics tools and their versions.
All tables (Supplementary files 1-5) are available as supporting online material. All of our own sequences, scripts and R workspace images are available on Github including the supplementary

files (github.com/UCanCompBio/Avoidance). The other datasets are cited in the manuscript (Supplementary file 3).

Evolutionary conservation

If excessive interactions between messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs) are detrimental to cellular function, then we expect the signature of selection against interactions (avoidance) to be a conserved feature of prokaryotic genomes. In the following, we describe where the data used to to test the evolutionary conservation of avoidance was acquisitioned, the models that we use to test avoidance and the negative controls in detail for evolutionary conservation predictions. We also investigate detect regions of avoidance on one of the core ncRNAs, the ribosomal small subunit (SSU) RNA.

Data sources for bacterial genomes

The bacterial genomes and annotations that we used for investigating mRNA:ncRNA interactions were acquired from the EBI nucleotide archive (2,564 sequenced bacterial genomes available on August 2013) (http://www.ebi.ac.uk/genomes/bacteria.html). We selected an evolutionarily conserved (core) group of 114 mRNAs from PhyEco (Wu et al. 2013) and an evolutionarily conserved (core) group of ncRNAs (Hoeppner et al. 2012). PhyEco markers are based on a set of profile HMMs that correspond to highly conserved bacterial protein coding genes (these include ribosomal proteins, tRNA synthetases as well as other components of translation machinery, DNA repair and polymerases) (Wu et al. 2013). The HMMer package (version 3.1b1) (Eddy 2011) was used to extract the mRNAs corresponding to these marker genes from genome files. We removed genome sequences that host fewer than 90% of the marker genes; leaving 1,582 bacterial genome sequences and 176,704 core mRNAs that spanned these.

We extracted the 1st to the 21st nucleotide of the core mRNAs. As this region showed the strongest signal in a small-scale analysis (Figure 3-figure supplement 1A), this region has also been shown to have an unusual codon distribution in previous work (Tuller & Zur 2015; Goodman et al. 2013) as explained in the main text.

We obtained ncRNA annotations using the Rfam database (version 11.0)(Gardner et al. 2011) for the well conserved and highly expressed tRNA, rRNA, RNase P RNA, SRP RNA, tmRNA and 6S RNA families (Rfam accessions: RF00001, RF00005, RF00010, RF00011, RF00013, RF00023, RF00169, RF01854, RF00177). The redundant annotations were filtered for overlapping and identical paralogous sequences, leaving 99,281 core ncRNA that spanned 1,582 bacterial genomes.

Data sources for archaeal genomes

We followed a similar pipeline for archaeal genomes as described for bacterial genomes. In total we processed 240 archaeal genomes, and after filtering those that had fewer than 90% of the marker genes, we had 118 archaeal genomes for further analysis (genomes available on August 2013) (http://www.ebi.ac.uk/genomes/archaea.html). These genomes host 12,370 and 10,804 core mRNAs and core ncRNAs respectively.

Test of an (extrinsic) avoidance model

We used RNAup (version 2.0.7)(Lorenz et al. 2011) to calculate the binding minimum (Gibbs) free energy (MFE) values of mRNA:ncRNA interactions. The RNAup algorithm combines the intramolecular energy necessary to open binding sites with intermolecular energy gained from hybridization (Mückstein et al. 2006). In other words, this approach minimizes the sum of opening intramolecular energies and the intermolecular energy (Figure 1-figure supplement 1C). In our model of avoidance, we test for a reduction in absolute binding MFE relative to negative controls as a measure of avoidance.

After testing a variety of negative controls (e.g. dinucleotide preserved shuffled mRNAs, the 5` end of homologous mRNAs from a different bacterial phylum, 100 nucleotides downstream of designated interaction region, reverse complements, and identically sized intergenic regions), we selected the dinucleotide frequency preserved shuffled sequences as our negative control since this displayed the most conservative interaction MFE distribution (Figure 3-figure supplement 1A,B,C). In more detail, to serve as a negative control we compute the interaction MFE between each of the core ncRNAs and 200 dinucleotide-preserved shuffled versions of the 5` end

mRNAs. A dinucleotide frequency preserving shuffling procedure is used as Gibbs free energies are computed over base pair stacks, i.e. a dinucleotide alphabet, therefore this method has been shown to be important in order to minimise incorrect conclusions (Workman & Krogh 1999). We tested if the energy difference between native and shuffled interaction distributions is statistically significant using the nonparametric one-tailed Mann–Whitney U test, which returns a single *P* value per genome (Figure 1C). If the distribution of native interaction energies for a genome is significantly higher (i.e. fewer stable interactions) than the negative control, this is an indication that the genome has undergone selection for mRNA:ncRNA avoidance. To create the background density difference lines (seen in grey at Figure 1B), we randomly selected 100 bacterial strains and plot differences between the densities of shuffled interactions.

Test of an intrinsic avoidance model

The energy-based avoidance model that we defined above is opaque to cases of "intrinsic avoidance". These are where the intrinsic properties of mRNA and ncRNA sequences restrict their ability to interact. For an extreme example, if ncRNAs are composed entirely of guanine and cytosine nucleotides, whilst mRNAs are composed entirely of adenine and uracil nucleotides, then these will rarely interact. Therefore, our energy-based avoidance measures for native and shuffled interactions will both be near zero, and thus will not detect a significant energy shift between the native and control sequences. In order to account for some of these issues, we compared the G+C difference between core ncRNAs and core mRNAs. We used a nonparametric two-tailed Mann-Whitney U test to determine if there is a statistically significant G+C difference between the two samples: G+C of ncRNAs vs G+C of 5` end mRNAs (Figure 1D, E).

Sliding window analysis to detect regions of significance for avoidance on SSU ribosomal RNA

We hypothesise that heterogeneous signals of avoidance within ncRNA sequences may correspond to the accessibility of different ncRNA regions. For example, are highly avoided regions of abundant ncRNAs more accessible than those that are avoided less? To create an avoidance profile, we tested binding MFEs of native and shuffled interactions throughout the

full-length SSU ribosomal RNA of *T. thermophilus*, using a one tailed Mann-Whitney U tests to evaluate the degree of avoidance for each nucleotide in the SSU rRNA (Figure 3) with a windows size of 10 and step size of 1 (Supplementary file 4). We selected the protein data bank (PDB) entry (4WZO) as it is one of the few ribosomal structures with associated protein, mRNA, tRNA and LSU binding data (Rozov et al. 2015). The native interactions are the interactions between *T. thermophilus* core mRNAs and SSU ribosomal RNA. The shuffled controls are derived from 200 dinucleotide preserved shuffled versions of the RNAs. We created a 2x2 contingency table which separates the counts of residues that either host a strong avoidance signal or little avoidance signal (regions with $P < 0.001$, Mann-Whitney U test) and residues that we predict to either be in contact (<3.4 Angstroms between atoms) with ribosomal proteins or ribosomal, transfer or messenger RNAs or not in contact with other molecules (i.e. accessible) (Figure 3). We applied a Fisher's exact test (Fisher 1922) to these groups to and discovered a statistically significant relationship between avoidance and accessibility ($P = 2.5\text{x}10^{-17}$).

We have applied the same analysis to the other *T. thermophilus* core ncRNA genes (tRNAs, tmRNA, RNase P RNA and SRP RNA) in order to determine regions of avoidance (Figure 3-figure supplement 3). Since there are more than one tRNAs, we aligned the cellular RNAs to the associated Rfam model (RF00005) (Gardner et al. 2011; Nawrocki et al. 2015) using the *cmalign* tool (Nawrocki & Eddy 2013).

Proteomics/Transcriptomics & GFP expression

We predict that mRNAs with low avoidance values will produce fewer proteins for each mRNA transcript than those with high avoidance. In order to test this, we conducted a meta-analysis of proteomics and transcriptomics data and the relationship between this data and measures of mRNA and ncRNA avoidance. In the following section we describe the origins of the data we have used and the statistical analysis we use to test whether avoidance influences gene expression.

Data sources and statistics for mRNA, protein abundance & GFP expression

We compiled our data from five protein and mRNA quantification datasets, which consist of three *E. coli* (Laurent et al. 2010; Goodman et al. 2013; Lu et al. 2007) and two *P. aeruginosa* (Laurent et al. 2010; Kwon et al. 2014) (Supplementary file 3). We calculated Spearman's correlation coefficients (and associated *P* values) among the protein abundances and 5` end secondary structure (measured by intermolecular MFE), codon bias (measured by codon adaptation index (CAI)) and avoidance (Figure 2A). We have created single and multiple regression models to determine the explained variances by these parameters (Figure 2-figure supplement 3 and Supplementary file 5). These models show that avoidance explains more variance on average than secondary structure or codon bias. Up to 70 percent of the variation in GFP expression can be explained by including all the parameters and mRNA abundances (Figure 2-figure supplement 3).

CAI metric defines how well mRNAs are optimised for codon bias (Sharp & Li 1987). The CAI values were determined based on codon distribution patterns acquired from the core protein coding genes of *E. coli* BL21(DE3) (Accession: AM946981.2) (Wu et al. 2013) using Biopython libraries (version 1.6) (Cock et al. 2009).

The folding MFE predicts how stable the secondary structure of an RNA can be. The folding MFEs of GFP mRNAs were calculated using the RNAfold algorithm (version 2.0.7) (Lorenz et al. 2011). We restricted folding energy to first 37 nucleotides because the most significant correlation was previously reported for this region (Kudla et al. 2009).

We acquired previously published GFP data, associated fluorescence values and mRNA quantifications (Kudla et al. 2009) via personal communication. Our avoidance model showed the highest and most significant correlation with GFP expression in that dataset ($R_s = 0.65$, $P = 1.69 \times 10^{-20}$) (Figure 2A and Figure 2-figure supplement 2D,E,F). 5` end secondary structure ($R_s = 0.62$, $P = 5.73 \times 10^{-18}$) correlates slightly less than avoidance, while CAI does not correlate significantly ($R_s = 0.02$, $P = 0.4$).

Sliding window analysis to detect regions of significance for avoidance on mRNAs

In order to identify a region of mRNA that is consistent and unique in the datasets that we

applied evolutionary and expression analyses to we created an avoidance profile from the previously published GFP mRNAs (Kudla et al. 2009). We calculated binding MFEs using a window size of 21 with a 1 nucleotide step size, and for each region we computed the associated Spearman's correlation coefficients with *P* values. This analysis revealed the significance of the first 21 nucleotides on expression, this is consistent with previous results that identify initiation as the rate limiting step for translation (Tuller & Zur 2015; Plotkin & Kudla 2011). It also revealed other statistically significant regions with high correlation correlation coefficient throughout the GFP mRNAs (Figure 3-figure supplement 1A).

mRNA design

We have shown that avoidance is a broadly evolutionary conserved phenomenon and that it is significantly correlated with protein abundance relative to mRNA abundance. We now wish to test if avoidance can be used to design mRNA sequences that modulate the abundance of corresponding protein in a predictable fashion. We use a set of GFP mRNA constructs that all maintain the same G+C content, codon adaptation index (CAI) and internal secondary structure but host either very high or very low avoidance values. This procedure was repeated for the CAI and internal secondary structure values while maintaining a constant avoidance. The resulting 13 constructs were synthesised, transformed and expressed by commercial services. In the following paragraphs we explained how we design our GFP constructs, the experimental set-up and statistical analyses.

Green fluorescence protein (GFP) mRNA design

We sampled 537,000 synonymous mRNA variants of a GFP mRNA (the 239 AA, 720 nucleotide long, with accession AHK23750, can be encoded by 7.62×10^{111} possible unique mRNA variants). In brief, these mRNA variants were scored based upon (1) CAI, (2) mRNA secondary structure in their 5` end region, and (3) mRNA:ncRNA interaction avoidance in their 5` end region.

The genome of *E. coli* BL21 encodes 52 unique core ncRNAs (Gardner et al. 2011; Nawrocki et al. 2015), to estimate the level of ncRNA avoidance for each GFP mRNA, we sum the binding

MFEs. For example, for each GFP mRNA we compute 52 independent binding MFE values for each ncRNA. In short, a higher summed MFE score for a GFP mRNA implies a higher avoidance, while a lower summed MFE score implies a lower avoidance. This approach assumes that the ncRNAs are expressed at much higher levels than GFP mRNAs (i.e. [ncRNA] >> [mRNA]) (Figure 4). Consequently, any potential interaction site on GFP mRNAs are likely to be saturated with ncRNA.

Finally, we selected 13 GFP mRNA constructs, while controlling the range of G+C values. These GFP mRNAs were designed to have four different aspects; extreme 5` end secondary structure (2 minimum and 2 maximum folding MFE constructs), extreme codon bias (2 maximum and 2 minimum CAI constructs), extreme interaction avoidance (2 minimum and 2 maximum binding MFE constructs) and an "optimal" construct. The optimal construct was selected for a high CAI, low 5` end structure and high avoidance. All extreme GFP mRNA constructs have near identical G+C content (between 0.468-0.480) and identical G+C contents at the 5` end (0.48). Each of the sampled GFP mRNAs is separated from other mRNAs by at least 112 nucleotide substitutions and 122 nucleotide substitutions on average (Figure 2-figure supplement 1).

Extreme GFP transformations, determining fluorescence levels and RT-qPCR analyses

Both GFP expression assays and RT-qPCR analyses were performed as part of a commercial service offered by the University of Queensland, Protein Expression Facility and Real-Time PCR Facility. Plasmid DNA from each construct was transformed into a expression strain of *E. coli* BL21(DE3). Starter cultures were grown in quadruplicate from single colonies in 0.5 mL of TB kanamycin 30 μg/mL media in a 96 deep-well microplate and incubated at 30°C, 400 rpm (3 mm shaking throw). Each starter culture was used to inoculate 1.0 mL of the same media at a ratio of 1:50, each in a single well of a 96 deep-well plate. The cultures were incubated at 30°C, 400 rpm for 1 hour, at this point the cultures were chilled for 5 min then induced into 0.2 mM IPTG and incubated at 20°C. For analysis, culture samples of 100 μL were taken at 1 hr, 2 hrs, 3 hrs, 4 hrs and 22 hrs (overnight) hours post-induction (HPI) for fluorescence and optical density analysis. Samples were collected in PetriWell 96-well flat bottom, black upper, lidded microplates

(Genetix). Cell density of fluorescence measurements were performed on a Spectramax M5 Microplate Reader using SMP software v 5.2 (Molecular Devices). For fluorescence intensity measurements, samples were collected in the 96-well plate listed above. Samples were analysed by bottom-read, 10 reads per well at an excitation wavelength = 488 nm, emission wavelength = 509 nm with an automatic cut-off at 495 nm and measured as relative fluorescence units (RFU). The raw RFU values were normalised by subtracting the averaged baseline values obtained from untransformed BL21(DE3) at the same time point. All samples at the 22 HPI time point were diluted 1:4 in TB kanamycin 30 μg/mL media before measurement.

Total RNA was purified from induced 0.5 mL of BL21(DE3) cultures on Maxwell® 16 robot (Promega) using LEV simplyRNA Tissue Kit (Promega). RNA concentrations were assessed on Qubit 3.0 Fluorometer (Thermo Fisher Scientific). cDNA synthesis was done using ProtoScript II First Strand cDNA Synthesis Kit (NEB) according to manufacturer protocol using random primer. The rpsL gene was selected as the reference gene (internal control). RT-qPCR was performed in 384-well plates with a ViiA™ 7 Real-Time PCR System (Thermo Fisher Scientific) using Life Technology SYBR Green-based PCR assay. The data analysis was performed using Applied Biosystems QuantStudio software (Thermo Fisher Scientific). The total volume of reaction was 10 μL including 0.2 μM of each primer as a final concentration. The following PCR conditions were used: 95ºC for 10 min, followed by 40 cycles of 95ºC for 15 s and 60ºC for 1 min. The melting curves were analyzed at 60-95ºC after 40 cycles. RNA concentrations were subsequently estimated using the $2^{-\Delta\Delta C_T}$ approach (Schmittgen & Livak 2008). We shared the raw data, oligos and primers in the supplementary files (Supplementary file 2A,B).

Statistical analyses of extreme GFP data

As described, we designed extreme GFP mRNA constructs, and measured the associated fluorescence. A Kruskal-Wallis test (nonparametric alternative of ANOVA) shows a statistically significant difference between the fluorescence of GFP mRNA groups ($P = 1.35\text{x}10^{-5}$) (Figure 2-figure supplement 1). Our pairwise comparison of GFP groups using a Kruskal-Nemenyi test (a nonparametric alternative of the Student's t-test) for fluorescence difference also reveals a

statistically significant difference in fluorescence between high avoidance constructs and low avoidance constructs ($P = 0.00036$).

We computed the Spearman's correlation coefficients (and associated *P* values) between GFP expression and each of the following measures; CAI ($R_s = 0.29$, $P = 0.016$), intramolecular folding energy ($R_s = 0.34$, $P = 0.006$), avoidance (intermolecular binding energy) ($R_s = 0.56$, $P = 6.9 \times 10^{-6}$) and mRNA concentration ($R_s = 0.73$, $P = 3.2 \times 10^{-3}$) to predict effect size of each predictor. Our avoidance model resulted in the highest correlation with GFP expression (Figure 2B,C,D).

Acknowledgements

SUU is supported by a Biomolecular Interaction Centre and UC HPC (Bluefern) joint PhD Scholarship from the University of Canterbury. AMP & PPG are both supported by Rutherford Discovery Fellowships, administered by the Royal Society of New Zealand. RCJD acknowledges the Royal Society of New Zealand Marsden Fund and US Army Research Office for funding support. Thanks to Grzegorz Kudla for sharing the GFP expression data from Kudla *et al* (2009) and Cindy Chang, Emilyn Tan, Michael Nefedov from the RT-PCR and the PEF facilities at the University of Queensland for assistance with generating GFP expression data. We also acknowledge Jeppe Vinther, Lukasz Kielpinski, Anders Krogh and the attendees of the 2012 and 2015 Benasque RNA conference for stimulating discussions.

References

Andersson, S.G. & Kurland, C.G., 1990. Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2), pp.198–210.

Bartel, D.P. & Chen, C.Z., 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature reviews. Genetics*, 5, pp.396–400.

Bhaya, D., Davison, M. & Barrangou, R., 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, pp.273–297.

Boël, G. et al., 2016. Codon influence on protein expression in E. coli correlates with mRNA

levels. *Nature*, 529(7586), pp.358–363.

Borg, A. & Ehrenberg, M., 2015. Determinants of the rate of mRNA translocation in bacterial protein synthesis. *Journal of molecular biology*, 427(9), pp.1835–1847.

Chamary, J.V. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9), p.R75.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Cock, P.J.A. et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* , 25(11), pp.1422–1423.

Deutscher, M.P., 2006. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic acids research*, 34(2), pp.659–666.

Diwan, G.D. & Agashe, D., 2016. The Frequency of Internal Shine-Dalgarno-like Motifs in Prokaryotes. *Genome biology and evolution*, 8(6), pp.1722–1733.

van Dongen, S., Abreu-Goodger, C. & Enright, A.J., 2008. Detecting microRNA binding and siRNA off-target effects from expression data. *Nature methods*, 5(12), pp.1023–1025.

Eddy, S.R., 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10), p.e1002195.

Farh, K.K.-H. et al., 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755), pp.1817–1821.

Fisher, R.A., 1922. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1), pp.87–94.

Gardner, P.P. et al., 2011. Rfam: Wikipedia, clans and the “decimal” release. *Nucleic acids research*, 39(Database issue), pp.D141–5.

Giannoukos, G. et al., 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome biology*, 13(3), p.R23.

Goodman, D.B., Church, G.M. & Kosuri, S., 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(6157), pp.475–479.

Hoeppner, M.P., Gardner, P.P. & Poole, A.M., 2012. Comparative Analysis of RNA Families Reveals Distinct Repertoires for Each Domain of Life C. O. Wilke, ed. *PLoS computational biology*, 8(11), p.e1002752.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the

occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology*, 151(3), pp.389–409.

Ioannidis, J.P.A. & Khoury, M.J., 2011. Improving validation practices in "omics" research. *Science*, 334(6060), pp.1230–1232.

Kudla, G. et al., 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* , 324(5924), pp.255–258.

Kwon, T. et al., 2014. Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. *Journal of proteome research*, 13(5), pp.2370–2380.

Lao, P.J. & Forsdyke, D.R., 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome research*, 10(2), pp.228–236.

Laurent, J.M. et al., 2010. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics*, 10(23), pp.4209–4212.

Li, G.-W., Oh, E. & Weissman, J.S., 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484(7395), pp.538–541.

Lindgreen, S. et al., 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS computational biology*, 10(10), p.e1003907.

Lorenz, R. et al., 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology: AMB*, 6, p.26.

Lu, P. et al., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25(1), pp.117–124.

Maier, T. et al., 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular systems biology*, 7, p.511.

Mückstein, U. et al., 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics* , 22(10), pp.1177–1182.

Nakahigashi, K. et al., 2014. Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC genomics*, 15, p.1115.

Nawrocki, E.P. et al., 2015. Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, 43(Database issue), pp.D130–7.

Nawrocki, E.P. & Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , 29(22), pp.2933–2935.

Pain, A. et al., 2015. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5), pp.509–513.

Pelletier, J. & Sonenberg, N., 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 65(6), pp.576–581.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1), pp.32–42.

Rozov, A. et al., 2015. Structural insights into the translational infidelity mechanism. *Nature communications*, 6, p.7251.

Schmittgen, T.D. & Livak, K.J., 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nature protocols*, 3(6), pp.1101–1108.

Sharma, C.M. et al., 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286), pp.250–255.

Sharp, P.M. & Li, W.-H., 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3), pp.1281–1295.

de Sousa Abreu, R. et al., 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12), pp.1512–1526.

Stark, A. et al., 2005. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6), pp.1133–1146.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.

Tuller, T. & Zur, H., 2015. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic acids research*, 43(1), pp.13–28.

Vivancos, A.P. et al., 2010. Strand-specific deep sequencing of the transcriptome. *Genome research*, 20(7), pp.989–999.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

Woolstenhulme, C.J. et al., 2015. High-precision analysis of translational pausing by ribosome

profiling in bacteria lacking EFP. *Cell reports*, 11(1), pp.13–21.

Workman, C. & Krogh, A., 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24), pp.4816–4822.

Wu, D., Jospin, G. & Eisen, J.A., 2013. Systematic identification of gene families for use as markers for phylogenetic and phylogeny- driven ecological studies of bacteria and archaea and their major subgroups. *PLoS ONE*, 8(10), p.e77033.

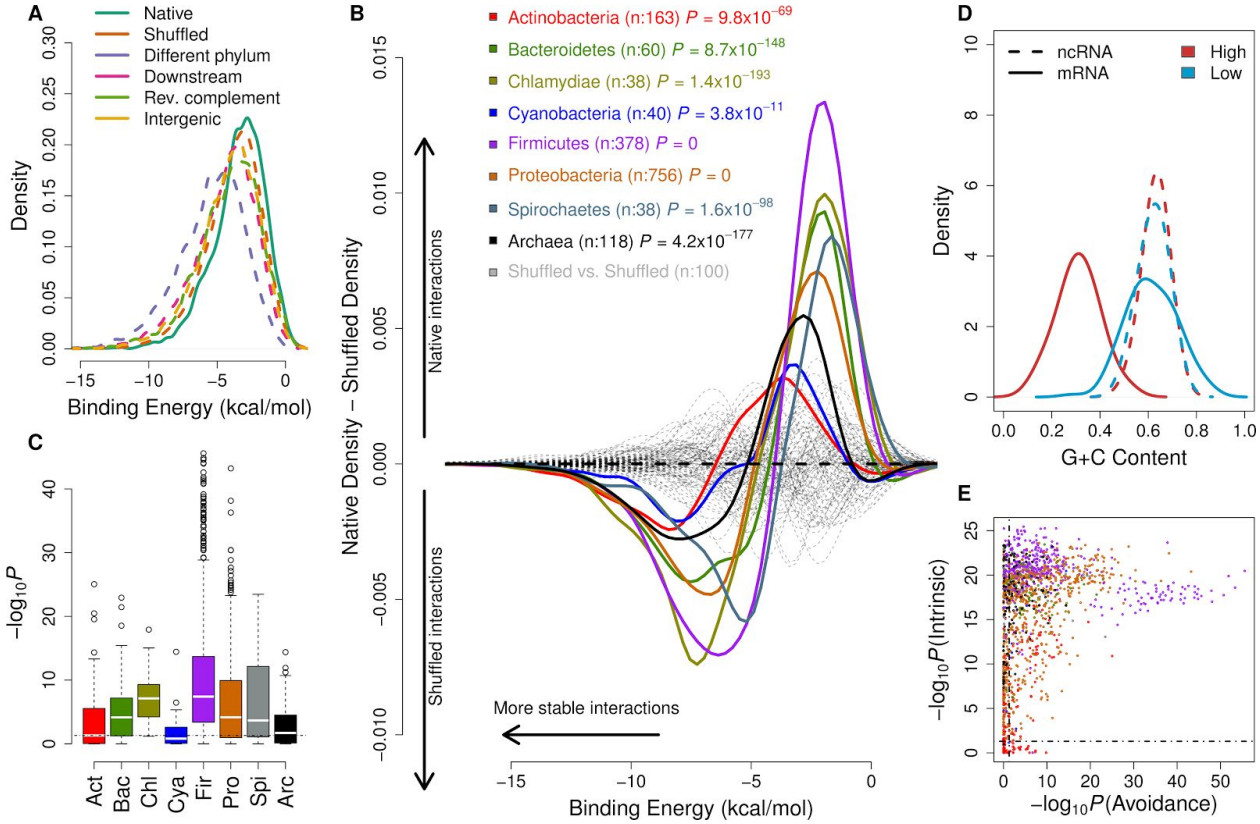

Figure 1. mRNA:ncRNA avoidance is a conserved feature of bacteria and archaea. **(A)** Native core mRNA:ncRNA binding energies (green line; mean = -3.21 kcal/mol) are significantly higher than all mRNA negative control binding energies (dashed lines; mean binding energies are -3.62, -5.21, -4.13, -3.86 & -3.92 kcal/mol respectively) in pairwise comparisons ($P < 2.2\text{x}10^{-16}$ for all pairs, one-tailed Mann–Whitney U test) for *Streptococcus suis* RNAs. **(B)** The difference between the density distributions of native mRNA:ncRNA binding energies and dinucleotide preserved shuffled mRNA:ncRNA controls as a function of binding energy for different taxonomic phyla. Each coloured curve illustrates the degree of extrinsic avoidance for different bacterial phyla or the archaea. Positive differences indicate an excess in native binding for that energy value, negative differences indicate an excess of interactions in the shuffled controls. The dashed black line shows the expected result if no difference exists between these distributions and the dashed grey lines show empirical differences for shuffled vs shuffled densities from 100 randomly selected bacterial strains. **(C)** This box and whisker plot shows $-\log_{10}(P)$ distributions for each phylum and the archaea, the P-values are derived from a one-tailed Mann–Whitney U test for each genome of native mRNA:ncRNA versus shuffled mRNA:ncRNA binding energies. The black dashed line indicates the significance threshold ($P < 0.05$). **(D)** A high intrinsic avoidance strain (*Thermodesulfobacterium sp.* OPB45) shows a clear separation between the G+C distribution of mRNAs and ncRNAs ($P = 9.2\text{x}10^{-25}$, two-tailed Mann-Whitney U test), and a low intrinsic avoidance strain (*Mycobacterium sp.* JDM601) has no G+C difference between mRNAs and ncRNAs ($P = 0.54$, two-tailed Mann-Whitney U test). **(E)** The x-axis shows $-\log_{10}(P)$ for our test of extrinsic avoidance using binding energy estimates for both native and shuffled controls, while the y-axis shows $-\log_{10}(P)$ for our intrinsic test of avoidance based upon the difference in G+C contents of ncRNAs and mRNAs. Two perpendicular dashed black lines show the threshold of significance for both avoidance metrics. 97% of bacteria and archaea are significant for at least one of these tests of avoidance.

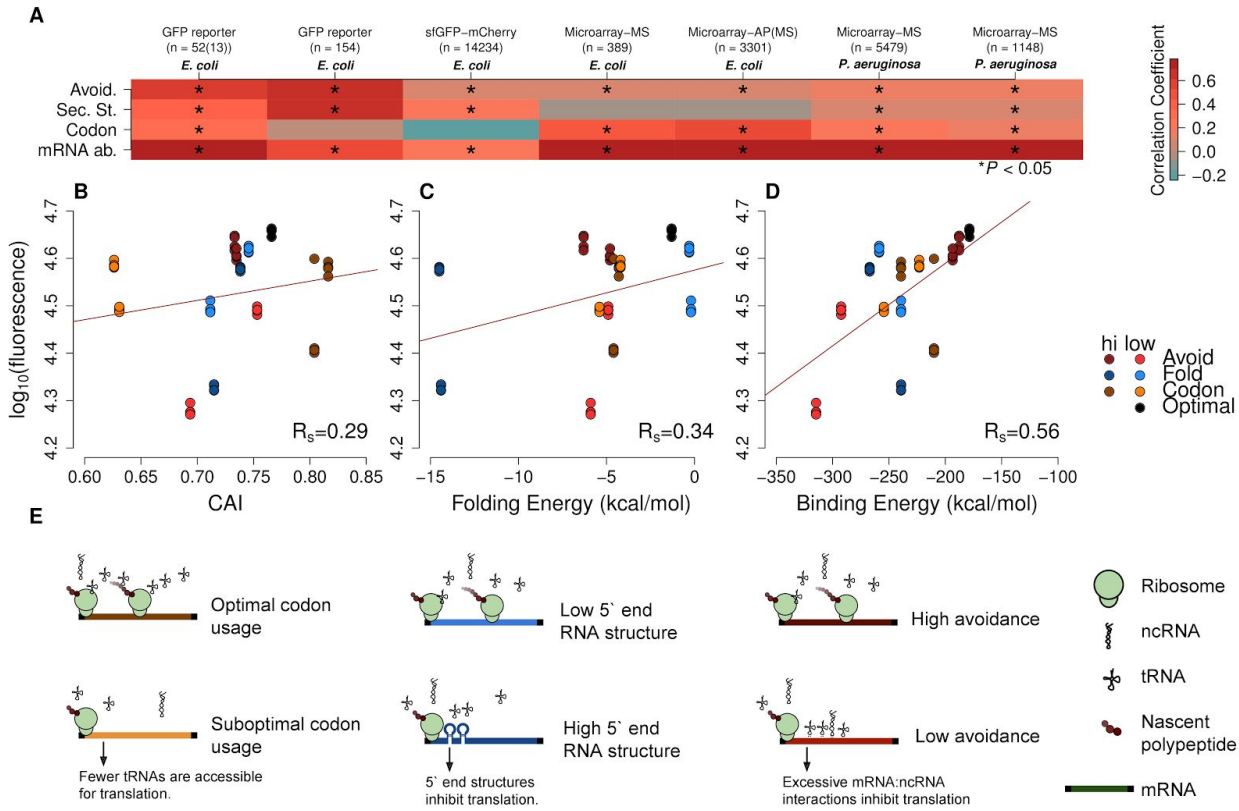

Figure 2. mRNA attributes have different impacts on protein abundance. **(A)** This heatmap summarizes the effect sizes of four mRNA attributes (avoidance of mRNA:ncRNA interaction, 5' end secondary structure, codon bias and mRNA abundance) on protein expression as Spearman's correlation coefficients, which are represented in gradient colors, while a starred block shows if the associated correlation is significant ($P < 0.05$). **(B)** GFP expression correlates with optimized codon selection, measured by CAI ($R_s = 0.29$, $P = 0.016$). **(C)** GFP expression correlates with 5` end secondary structure of mRNAs, measured by 5' end intramolecular folding energy ($R_s = 0.34$, $P = 0.006$). **(D)** GFP expression correlates with avoidance, measured by mRNA:ncRNA binding energy ($R_s = 0.56$, $P = 6.9 \times 10^{-6}$). **(E)** Each cartoon illustrates the corresponding hypothesis; (1) optimal codon distribution (corresponding tRNAs are available for translation), (2) low 5` end RNA structure (high folding energy of 5` end) and (3) avoidance (fewer crosstalk interactions) lead to faster translation.

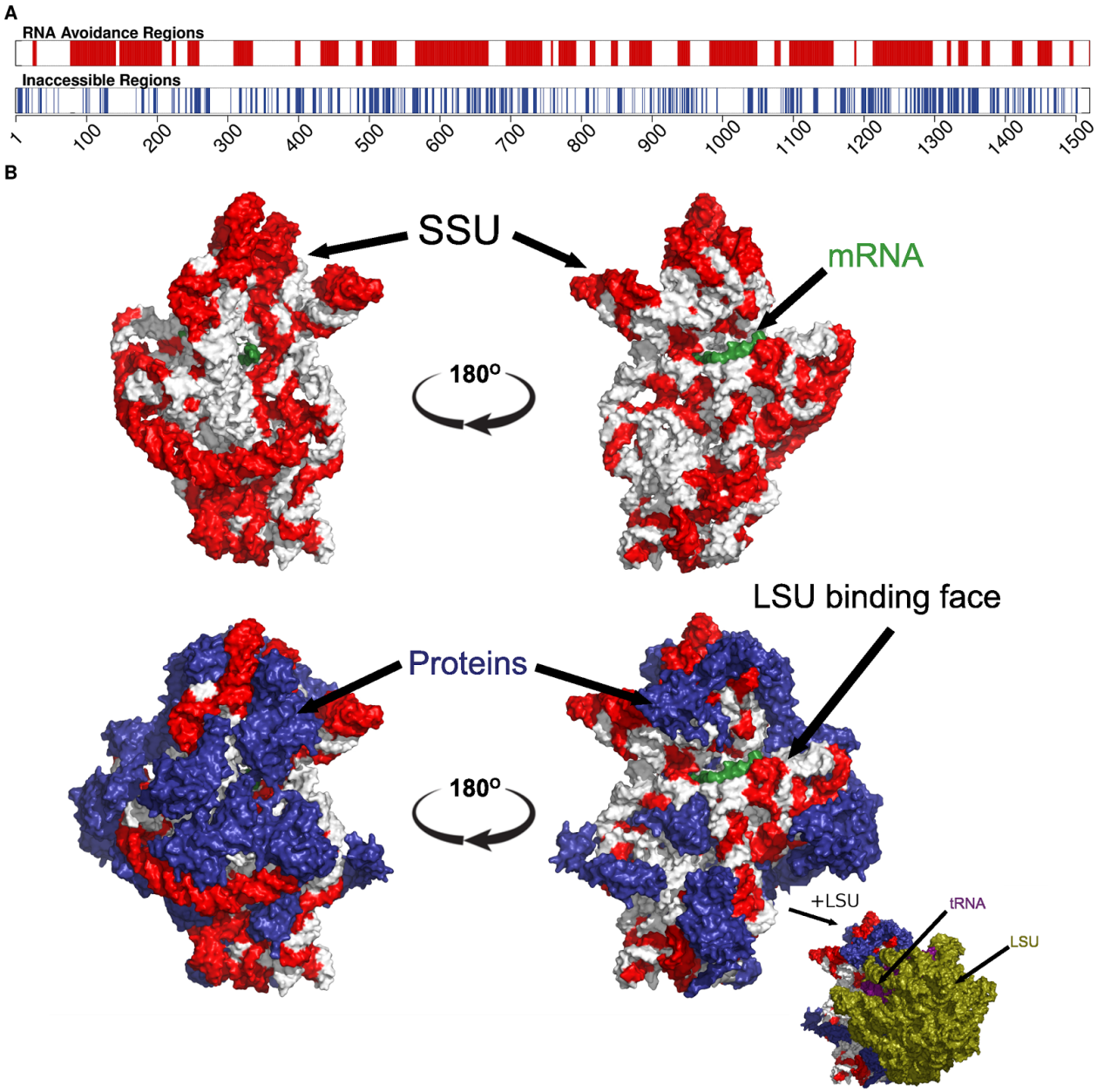

Figure 3. The most under-represented mRNA:rRNA interactions correspond to exterior regions of the ribosome. **(A)** In the upper bar, the regions of the *T. thermophilus* SSU rRNA that are under-represented in stable interactions with mRNAs ($P<0.05$) are highlighted in red. In the lower bar, the inaccessible residues (<3.4 Angstroms from other nucleotides or amino acids in the PDB structure 4WZO). **(B)** The 3 dimensional structure of the *T. thermophilus* ribosome includes 5S, SSU and LSU rRNA, 48 ribosomal proteins, 4 tRNA and a bound mRNA (PDB ID:

4WZO) (Rozov et al. 2015). We have highlighted the most avoided regions of the SSU rRNA in red (based upon the fewest stable interactions with *T. thermophilus* mRNAs ($P<0.05$). Two different orientations are shown on the left and right, the upper structure shows just the SSU rRNA and mRNA structures, the lower includes the ribosomal proteins (coloured blue). Bottom right, a view of the ribosome that also includes the LSU rRNA (green) is also shown. There is a significant correspondence between the accessibility of a region of SSU rRNA and the degree to which it is avoided ($P=2.5x10^{-17}$, Fisher's exact test).

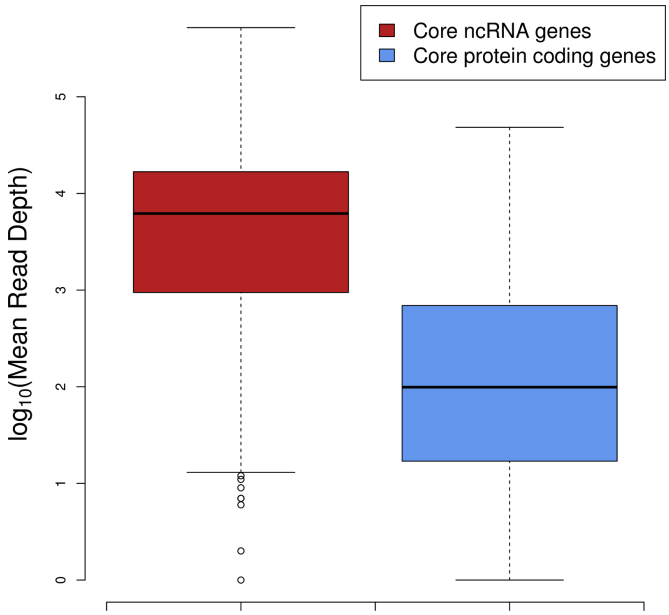

Figure 4. The median expression of core ncRNA genes (n=325 data points) in prokaryotic genomes is nearly two orders of magnitude greater than core mRNAs (n=8086 data points) which proves that ncRNAs constitute most of the cellular RNAs. To create this plot, we used mean mapped reads per gene length (i.e. mean read depth per position) of each core gene. The expression data is compiled from 5 archaeal and 37 bacterial strains from a previous study (Lindgreen et al. 2014).

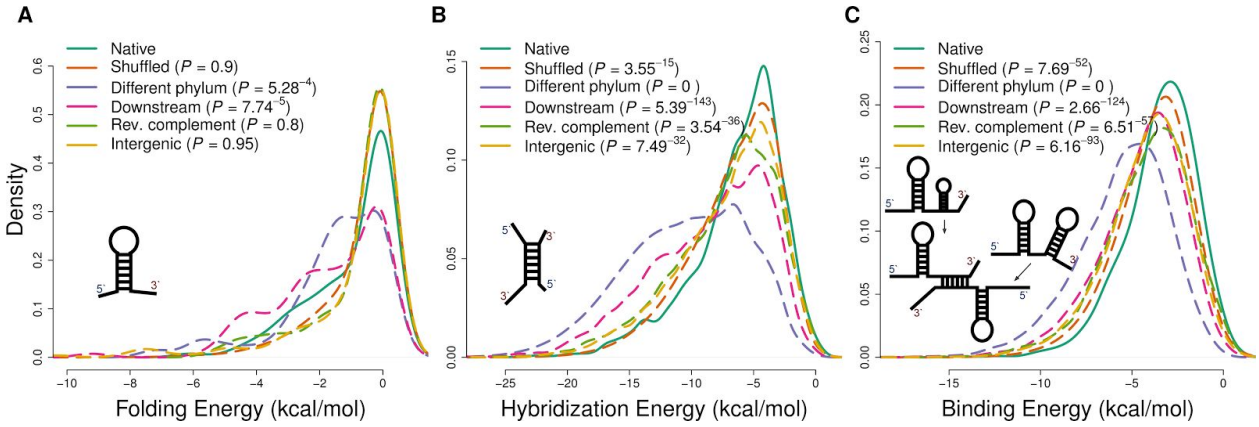

Figure 1-figure supplement 1. Applying different energy models of intramolecular and intermolecular interactions for native sequences and various negative controls. **(A)** The distributions of internal secondary structure (intramolecular) minimum free energies (MFEs) for 5` ends of mRNA sequences, estimated using RNAfold from the Vienna package (Lorenz et al. 2011). **(B)** The distributions of hybridization MFEs between core mRNAs and ncRNAs, estimated using the RNAduplex algorithm from the Vienna package (Lorenz et al. 2011). **(C)** The distributions of binding MFEs between core mRNAs and ncRNAs, estimated using the RNAup algorithm (Lorenz et al. 2011). The RNAup algorithm minimizes the sum of energies necessary to open binding sites on two RNA molecules and the hybridization energy (Lorenz et al. 2011). This method has been shown to be the most accurate general approach for sequence-based RNA interaction prediction (Pain et al. 2015).

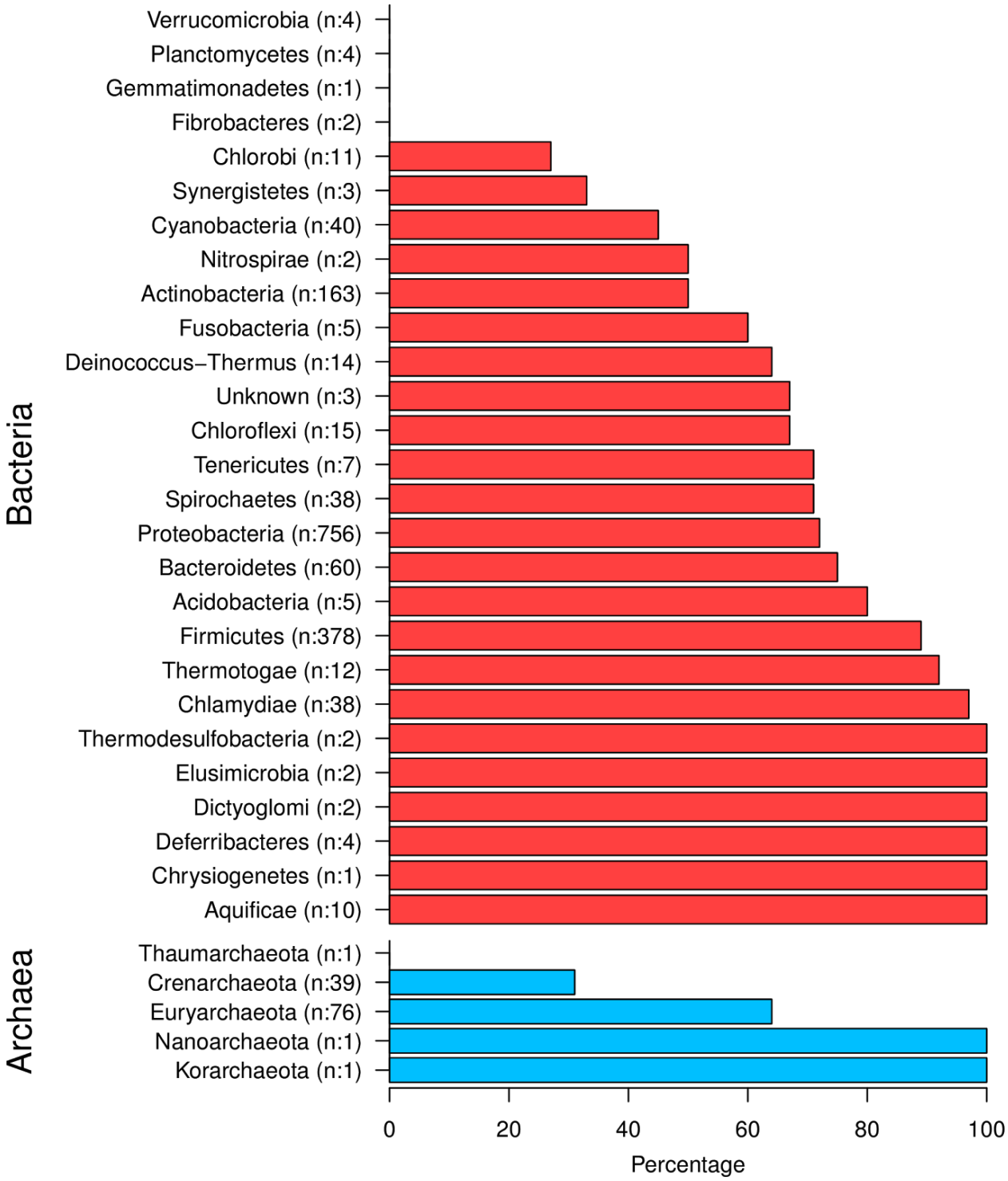

Figure 1-figure supplement 2. The top and the bottom panels shows bacterial phyla and archaeal phyla respectively. Numbers in brackets show the total members and the x-axis displays the percentage of extrinsic avoidance conservation in associated phylum. The archaeal and bacterial phyla with fewer than 20 publicly available sequenced genomes were excluded from

further analysis due to concerns about sample size sufficiency.

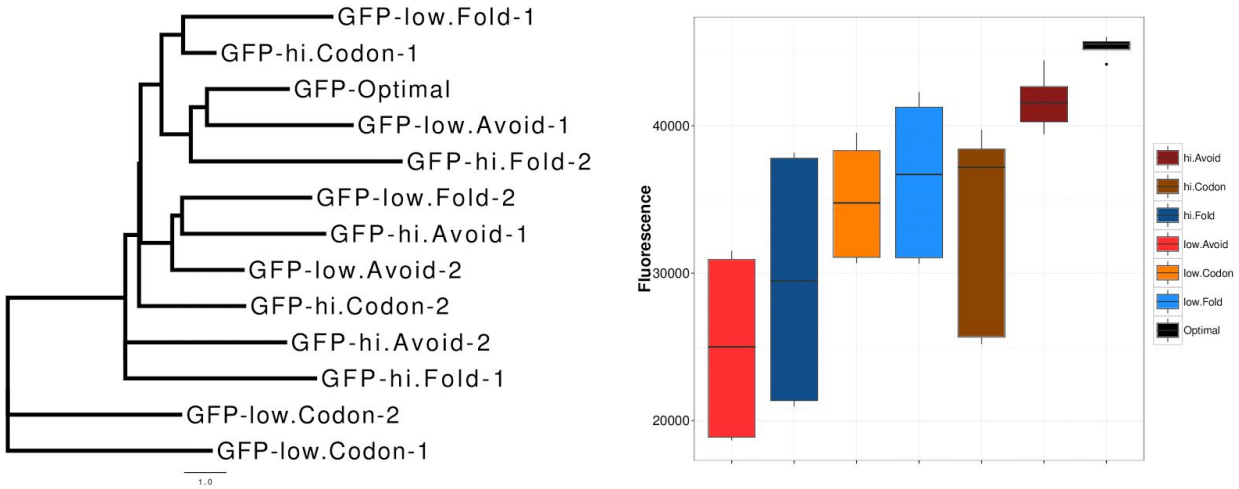

Figure 2-figure supplement 1. GFP mRNA constructs have unbiased design that produce different protein expressions. An unrooted maximum likelihood tree of the extreme GFP mRNAs on the left panel illustrates the low similarity between our GFP mRNA constructs. The distances were calculated using HKY85 nucleotide substitution model. On the right panel, the y-axis shows relative fluorescence units (RFU) of GFP expression from synonymously sampled mRNAs with different characteristics, these are labelled on the figure legend. Optimal and high avoidance GFP mRNAs produce the highest expression while low avoidance GFP mRNAs have the lowest expression ($P = 1.35\text{x}10^{-5}$, Kruskal-Wallis test).

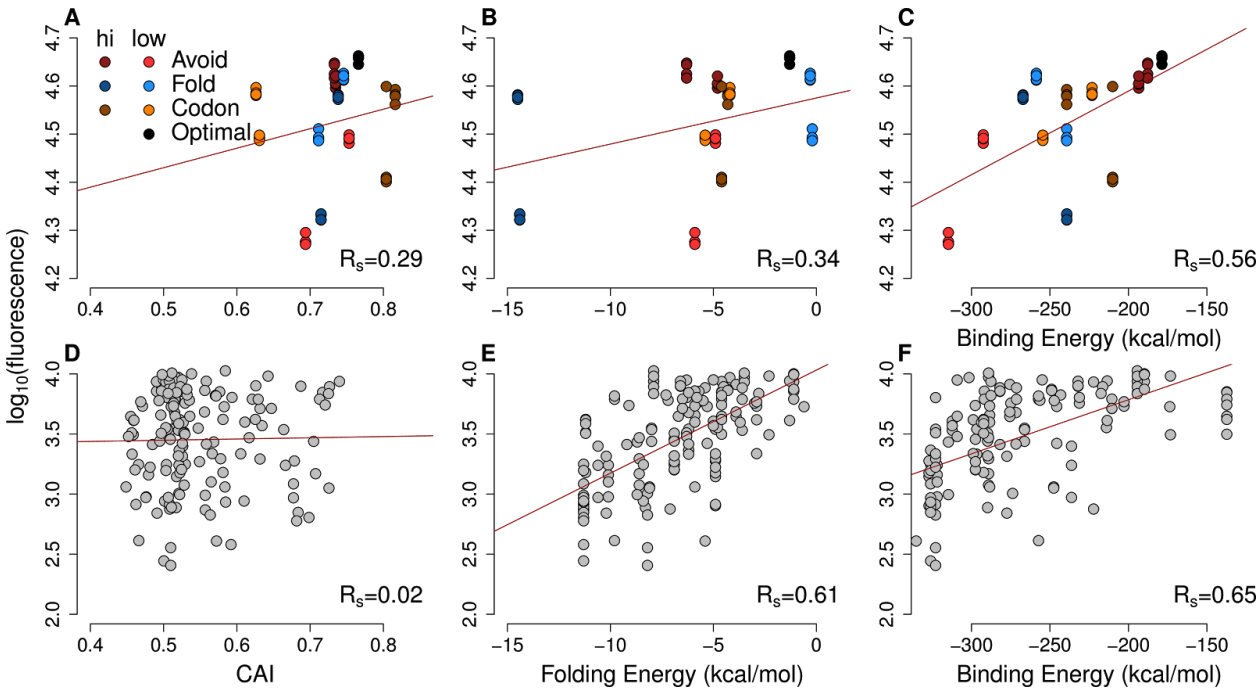

Figure 2-figure supplement 2. The scatter-plots of protein abundances (as log-fluorescences) summarize the effect of general factors for extreme GFP and previously published GFP datasets. **(A) (B) (C)** Each GFP mRNA was sampled from the extremes of one of three metrics presumed to impact expression mRNA:ncRNA binding, 5` end secondary structure or codon usage. Slightly darker or lighter colors display the type of extremes. Avoidance correlates with GFP expression ($R_s$ = 0.56, $P$ = 6.9x10$^{-6}$) more than CAI ($R_s$ = 0.29, $P$ = 0.01) and 5` end folding energy (R_s = 0.34, P = 0.006). **(D) (E) (F)** Using a previously published GFP dataset (Kudla et al. 2009) the CAI does not correlate with protein abundance (R_s = 0.02, P = 0.4), while 5` end folding energy (R_s = 0.61, P = 5.7x10^{-18}) and avoidance (R_s = 0.65, P = 1.6x10^{-20}) influence GFP expression.

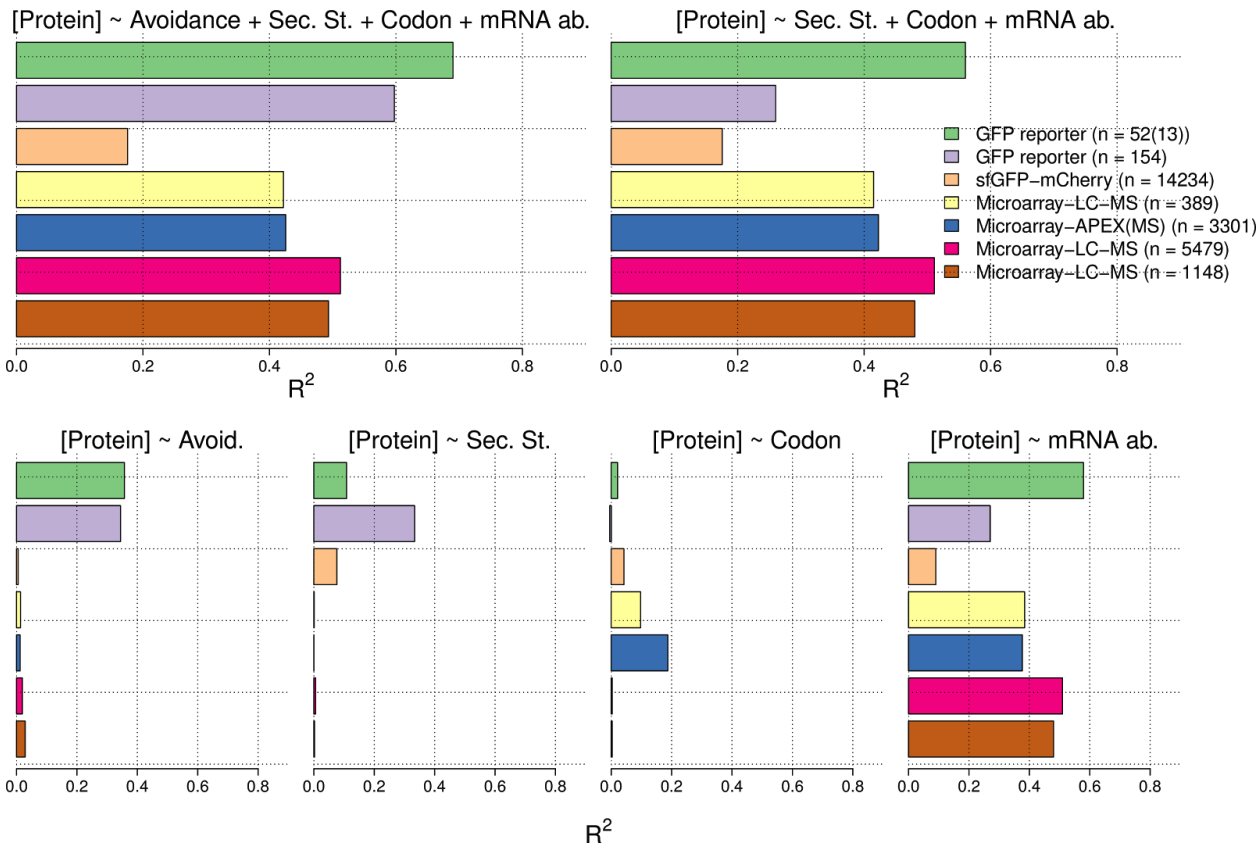

Figure 2-figure supplement 3. In the lower four panels we show the R^2 values for linear regression models between measures of each of avoidance, internal secondary structure, codon usage and mRNA levels for each of seven independent protein and mRNA expression datasets (Supplementary table 5). We have also computed R^2 values for multiple linear regression models of the sum of the four measures (right) and the sum less the avoidance measure (right).

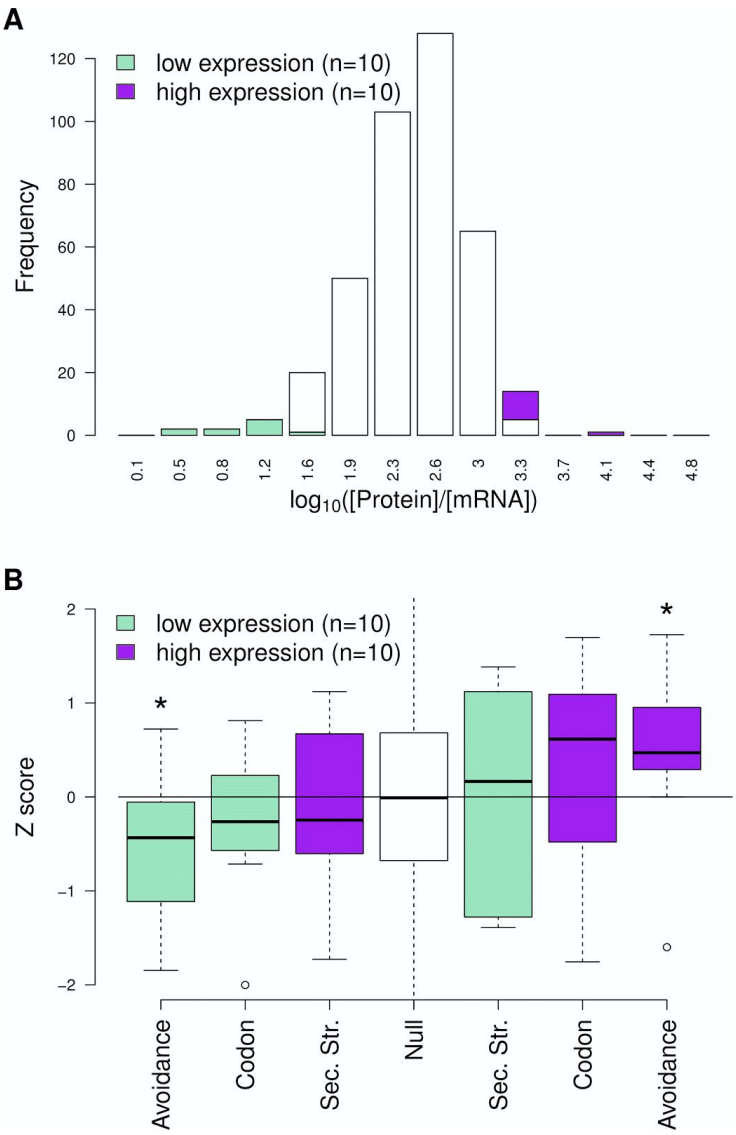

Figure 2-figure supplement 4. (A) In this plot a distribution of protein-per-mRNA ratio of native *E. coli* genes (n=389) (Laurent et al. 2010) is seen. We selected the top ten most and least productive genes which lie on the extreme ends of the plot (purple and green bars) **(B)** The y-axis shows the z-transformed scores of native mRNAs: CAIs, folding energies and binding energies. The expected background distribution (the white null bar in the middle) has a mean of 0 and standard deviation of 1, while a starred block shows whether the associated z-scores are significantly higher (or lower) than this background ($P < 0.05$). This demonstrates RNA avoidance is the only factor that explains protein-per-mRNA ratio difference of the most and the least efficient native *E. coli* mRNAs.

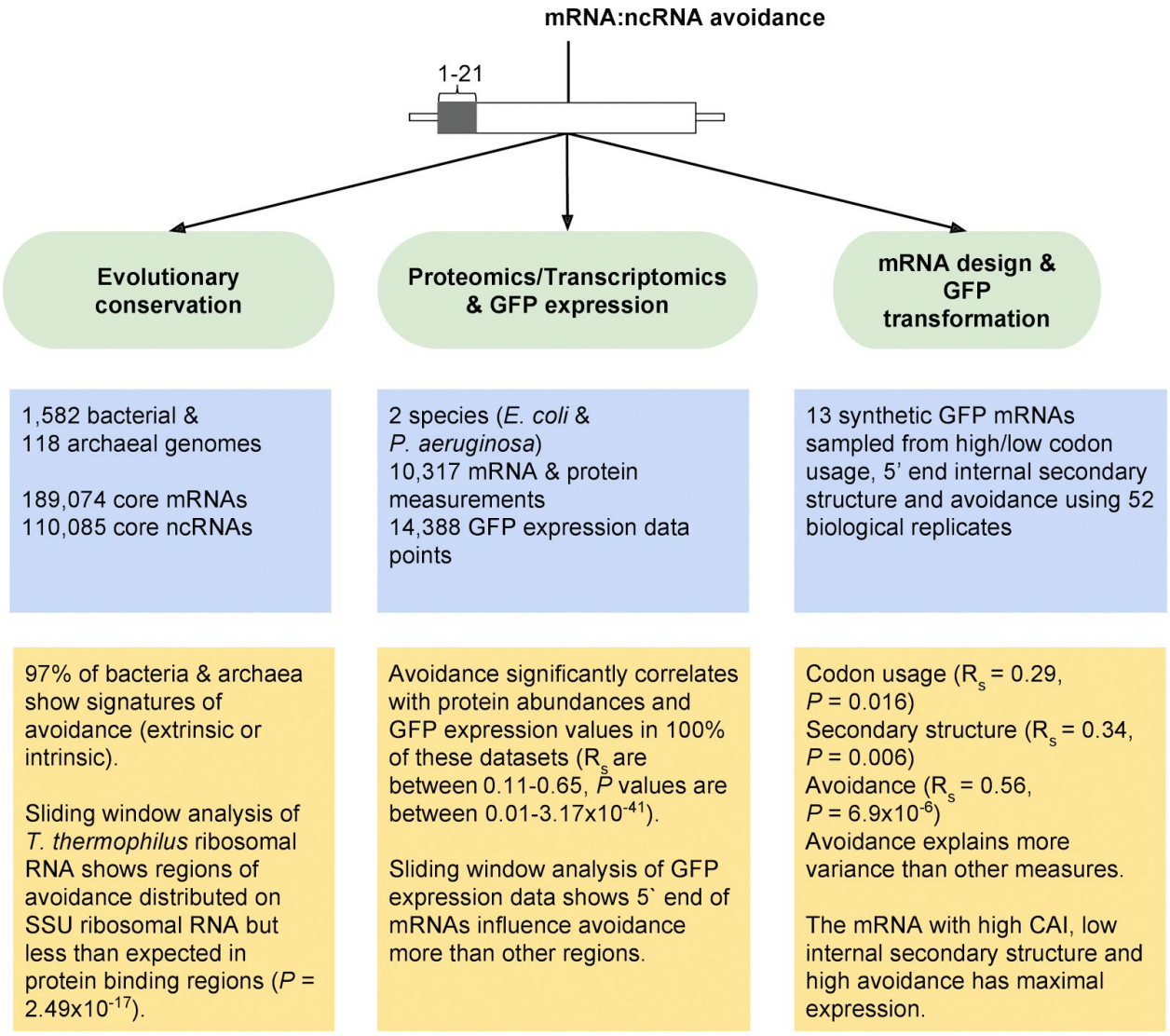

Figure 2-figure supplement 5. Overview of mRNA:ncRNA avoidance analysis and results. Our tests for avoidance can be divided into three main parts; (1) evolutionary conservation analyses to detect energy shifts in bacterial and archaeal genomes relative to dinucleotide shuffled negative controls, (2) analyses of proteomics, transcriptomics and GFP transformation data to predict the effect size of avoidance on protein expression and lastly (3) the application of avoidance hypothesis to design synonymous mRNAs that either produce high or low levels of corresponding protein.

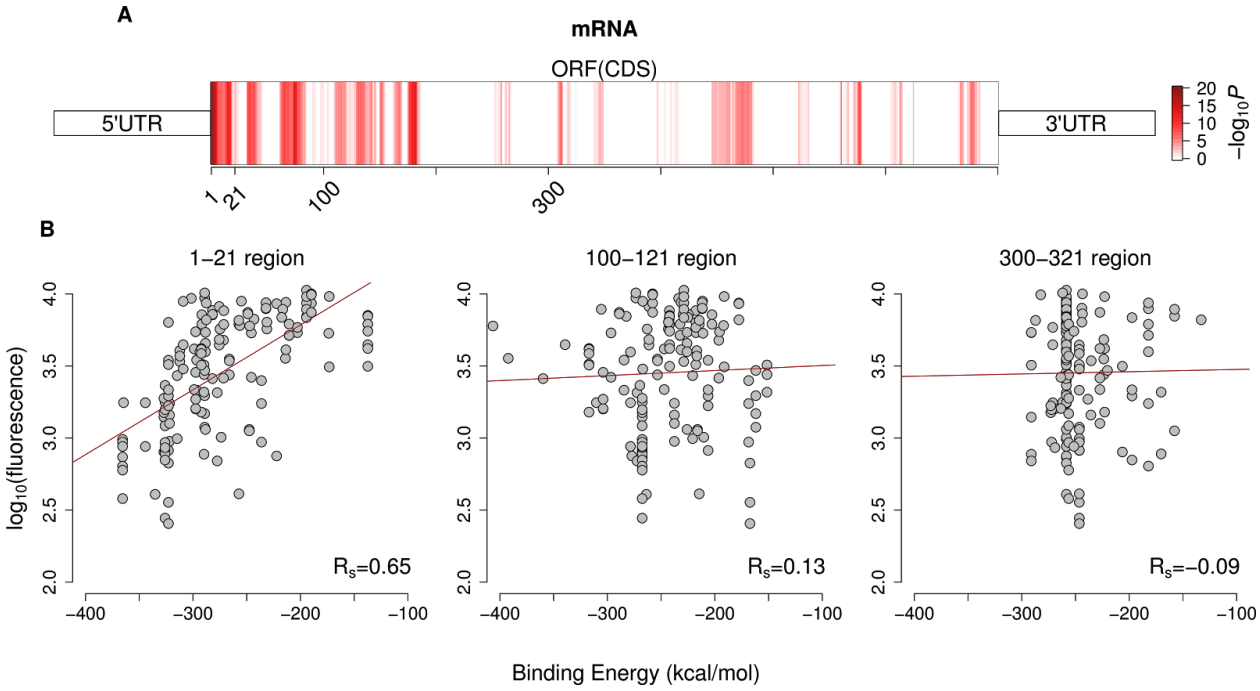

Figure 3-figure supplement 1. Avoidance pattern and its correlation with protein expression vary on mRNAs. **(A)** A sliding window (length 21, step size 1) analysis based on previously published GFP expression dataset (Kudla et al. 2009) shows the significance of correlation between avoidance and their corresponding fluorescence values for each position along the coding region. Darker red regions show more significant positions (with higher -$\log_{10}P$ values). **(B)** This analysis proves that binding energy of first 21 nt region influences protein expression more than any other downstream region and corresponding Spearman's correlation coefficients for selected sliding window start positions are seen at bottomright. It also justifies our selection of 5` end coding region for avoidance.

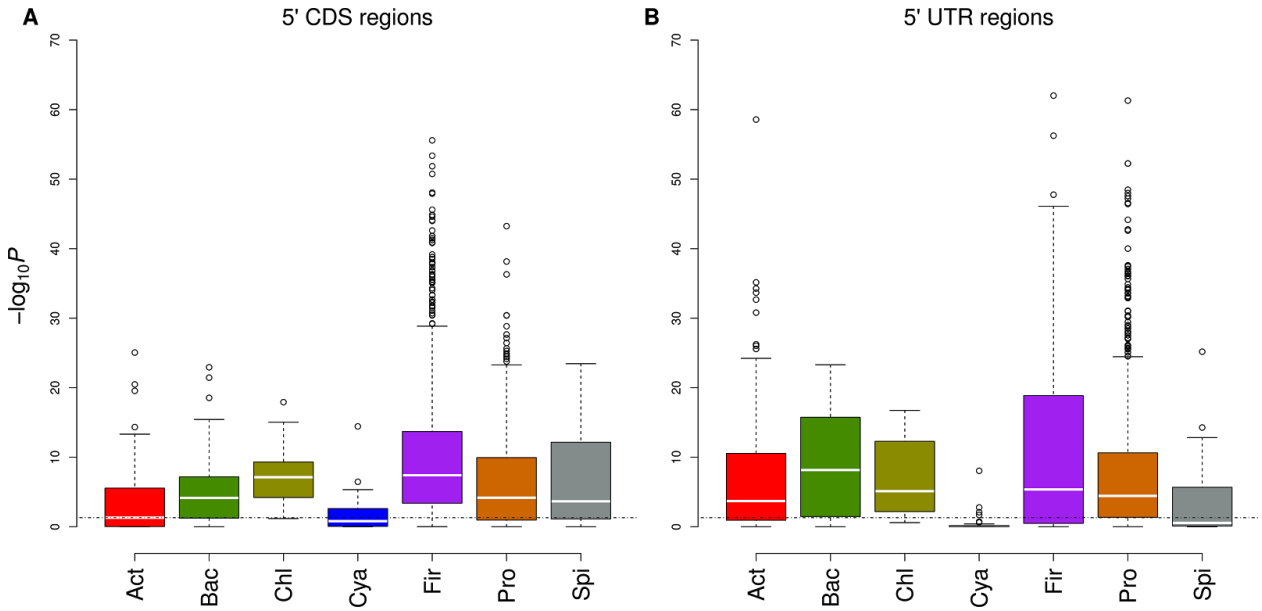

Figure 3-figure supplement 2. Comparison of different regions for evolutionary conservation analyses. **(A)** This box and whisker plot (similar with Figure 1C except archaea) shows -$\log_{10}(P)$ distributions for each bacterial phylum. The black dashed line indicates the significance threshold ($P < 0.05$). We used 5` end CDS regions as designated interaction location. **(B)** In this plot, 5` end UTR regions (90 nucleotides upstream to 21 nucleotides downstream) are used as designated interaction regions. It seems both regions have similar avoidance conservation, which proves avoidance is not limited to 5` ends of coding region.

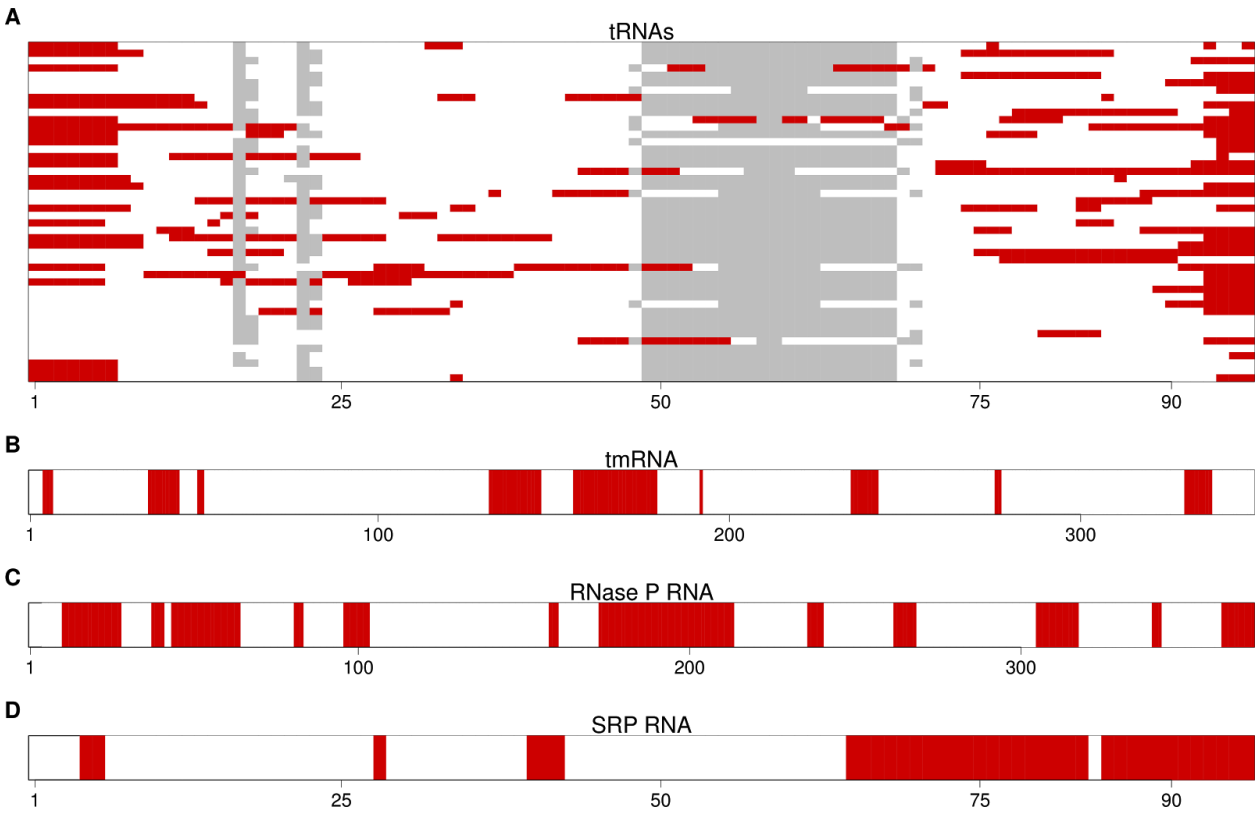

Figure 3-figure supplement 3. The most avoided regions of selected *T. thermophilus* non-coding RNAs. **(A)** A graphical view for an alignment of the *T. thermophilus* tRNAs (n=46). Regions that have significantly ($P < 0.001$, Mann-Whitney U test) fewer than expected interactions with *T. thermophilus* mRNAs are highlighted in red. These regions are therefore the most avoided regions by the host's mRNAs. The grey blocks show gaps in the alignment. **(B-D)** A graphical view of the most avoided regions is illustrated for tmRNA, RNase P and SRP RNA respectively.

CHAPTER V - Concluding Remarks

5.1 Conclusion

The primary aims of this thesis are to better understand the extent to which ncRNAs are coded in bacterial and archaeal genomes, and to examine the effects of crosstalk interactions in prokaryotes. Therefore, we focused on ncRNA discovery (Chapter II) and RNA-RNA interactions (the benchmark and RNA avoidance hypothesis in Chapter III, IV respectively) in archaea and bacteria (Figure 1.4).

5.1.1 Investigating Prokaryotic Dark Matter

Prokaryotic transcriptomes were considered to be relatively simple (Güell et al. 2009), but they recently have been shown to have transcriptional complexity comparable to eukaryotes (Barquist & Vogel 2015; Güell et al. 2009; Güell et al. 2011), and transcripts without any protein product are common in prokaryotes (Wade & Grainger 2014; Croucher & Thomson 2010; Güell et al. 2009; Chen et al. 2016; Sharma et al. 2010). However, there has been a lot of debate in the RNA community (for both eukaryotes and prokaryotes) about whether expression equals function, and what level of evidence is required to illuminate the 'genomic dark matter' (Palazzo & Gregory 2014; Wade & Grainger 2014; Pauli et al. 2015; Clark et al. 2011; van Bakel et al. 2010; Lloréns-Rico et al. 2016; Kapranov et al. 2007; Cohen et al. 2016).

First we asked the question 'can we identify novel ncRNA genes using available transcriptome data?'. We found evidence of nearly a thousand unnotated (most likely) ncRNA products, which were among the most abundant transcripts. In short, our 'An RNA Encyclopedia of Bacteria and Archaea' (AREBA) project results showed that (1) ncRNAs are highly abundant in prokaryotic genomes, (2) current RNA-seq data is heavily biased towards model organisms (or pathogens), and (3) most of the ncRNAs detected are evolutionarily young (Lindgreen et al. 2014). In this

project we processed over 400 publicly available archaeal and bacterial transcriptomics datasets (Lindgreen et al. 2014).

High-throughput methods (e.g. tiling-arrays, RNA-seq) usually produce thousands of candidate genes (i.e. potentially functional transcripts), which are hard to characterize as functional or nonfunctional. Since wet-lab experiments for verification (e.g. genetic screening) are time consuming, it is hard to examine all candidates one by one. Only using high expression as an indicator of function is also not a good idea, as some ncRNA genes are known for their very low expression (e.g. mammalian HOTTIP RNA ~ 0.3 transcripts per cell) (Clark et al. 2011; Wang et al. 2011).

Therefore, we have proposed a phylogeny-informed approach to classify novel ncRNA genes; otherwise, it is hard to characterize RNA transcripts generated by the high-throughput methods, and to separate them from transcriptional noise. In other words, RNA-seq strains must be selected inside an optimal phylogenetic zone that we refer to as 'the Goldilocks zone', because the phylogenetic window for the effective use of comparative methods for ncRNA identification is narrow. Otherwise, the signals of novel ncRNA genes could be lost due to sampling from too divergent or too close strains. Furthermore, over-sampling (i.e. creating redundant or unnecessary data) could waste valuable resources (e.g. sequencing cost, wet-lab cost and time etc.).

Our detailed computational investigation of the potentially functional ncRNA transcripts (nearly a thousand) based on their sequence conservation (e.g. HMMer package and iterative searching, reannotation, ORF prediction, RNAcode etc.) (Washietl et al. 2011; Wheeler & Eddy 2013; Mistry et al. 2013; Nawrocki & Eddy 2013), secondary structure conservation signals (e.g. RNAz) (Gruber et al. 2007), secondary structure stability (Lorenz et al. 2011), expression and phylogenetic distance (of the associated strains) produced a short list (i.e. 25 out of ~1000) of high-confidence expressed uncharacterized transcripts. We suggested that these short-listed

transcripts are ideal candidates for further verification, since they should contain fewer false positives overall.

Unfortunately, not many strains are inside the Goldilocks zone to allow for a better analysis on the available dataset as we mentioned (Chapter II). There are also other complementary bioinformatics methods that can be used on an ideal dataset such as RNA interaction prediction (Pain et al. 2015; Lai & Meyer 2015) for candidate sRNAs (i.e. test stability of interaction for candidate targets), differential expression analyses (i.e. comparing expression levels among stress conditions) (Love et al. 2014; McCarthy et al. 2012), functional enrichment (i.e. test for the possible functions) (Conesa & Götz 2008; Huang et al. 2009) and motif discovery (Gardner & Eldai 2015). We are planning to use these methods in the following study (i.e. AREBA-II).

In summary, our phylogeny-informed approach is an optimal way to design RNA-seq experiments for a better ncRNA detection. It provides the essential comparative information to characterize uncharacterized RNAs which can be used to evaluate genomic dark matter transcripts. Furthermore, the other high-throughput (i.e. CLIP-seq, ribosome profiling, Tn-seq etc.) and bioinformatics methods can be used in a combination with our method to increase the information harnessed (and to select strains for a transcriptome study).

5.1.2 A benchmark of RNA-RNA interactions

Many RNAs utilize RNA-RNA interactions to perform their roles, and computational methods are the most feasible alternative to interaction predictions and many algorithms have been developed to solve this problem. Some of them are ncRNA type specific (i.e. snoRNA and miRNA specialised tools), and some others try to predict all possible interactions for all domains of life (Chapter III). Therefore, we asked the question ‘which algorithm is the most successful one for all types of RNA-RNA interactions?’ The current RNA literature on RNA-RNA interaction prediction could not answer this yet. To answer this question, we benchmarked 15 different algorithms using dataset of 154 experimentally validated RNA-RNA interactions spanning all kingdoms of life: archaea, bacteria and eukaryota. This makes our benchmark one of

the most comprehensive ever published. Along with the current literature (Pain et al. 2015; Lai & Meyer 2015), our results showed that the energy based methods with accessibility are the most successful programs. As known, the accessibility methods calculate the energy needed to open the designated binding regions of interacting RNAs (Richter & Backofen 2012; Lorenz et al. 2011; Tafer & Hofacker 2008).

MFE methods are widely used for RNA interaction predictions, but they suffer from limited thermodynamic parameters and approximations (Gardner & Giegerich 2004; Dieterich & Stadler 2012; Layton & Bundschuh 2005; Mathews et al. 1999; Pain et al. 2015; Lai & Meyer 2015). Besides these drawbacks, one major problem of the current RNA-RNA interaction literature is the number of verified interactions. As we mentioned, many clades of life are not represented in available databases of verified interaction pairs (Wang et al. 2016; Chou et al. 2016). This causes overfitting of models to the available data.

To overcome this problem, we created one of the largest benchmark datasets in RNA interaction prediction literature, which contains verified interaction from all domains of life (Chapter III). Therefore, we tested the accuracy of algorithms on various datasets as much as possible. For example, IntaRNA was tested on bacterial sRNAs (Busch et al. 2008), but also appeared to successfully predict eukaryotic miRNA interactions as well, whereas RNAhybrid was developed for eukaryotic miRNAs (Krüger & Rehmsmeier 2006), but it is not very successful for bacterial sRNA predictions.

Creating dinucleotide shuffled RNA sequences (Workman & Krogh 1999) for use as negative controls is another way to detect stable interactions. This approach enables the creation of a background distribution of MFE values to test the statistical significance of a native interaction pair. However, we did not use this approach in our benchmark, as some algorithms do not produce MFE values (e.g. RactIP and bistaRNA) (Poolsap et al. 2011; Kato et al. 2010) or are biased towards internal structures (e.g. pairfold and RNAcofold) (Lorenz et al. 2011; Andronescu et al. 2003) that increase the false positive rate. We did however use this approach

for the interaction predictions of the RNA avoidance study, which helped us to reveal the MFE shifts for evolutionarily conservation analysis (Chapter IV - Materials and Methods).

In summary, our benchmark study provides us with the most accurate RNA-RNA interaction prediction method based on the available verified interactions. We used the selected algorithm for the predictions required in the RNA avoidance study. Yet, one must keep in mind that the current prediction algorithms are not perfect. They introduce a lot of false positives to their predictions, and occasionally miss true positives (Chapter III) (Pain et al. 2015).

5.1.3 Investigating crosstalk mRNA-ncRNA interactions in prokaryotes

Protein and mRNA levels are not strongly correlated (de Sousa Abreu et al. 2009; Vogel & Marcotte 2012; Kwon et al. 2014; Maier et al. 2011; Lu et al. 2007; Taniguchi et al. 2010; Chen et al. 2016). There are two well-known global factors that influence protein expression (Tuller et al. 2010), which explain some variation in mRNA and protein levels. These two factors are codon usage (or codon bias) (Tuller et al. 2010; Ikemura 1981; Ikemura 1985; Akashi 1994) and secondary structure of mRNAs (Chamary & Hurst 2005; Tuller et al. 2010; Gaspar et al. 2013; Pelletier & Sonenberg 1987; Gu et al. 2014). Yet, at best, these features account for only up to half of this variation (Kudla et al. 2009; Maier et al. 2011; Plotkin & Kudla 2011; Goodman et al. 2013; Chen et al. 2016).

Since ncRNAs are highly abundant in prokaryotic genomes (Chapter IV) (Lindgreen et al. 2014), we have proposed that they may have a ‘crosstalk’ effects on protein expression. Despite a growing recognition of the importance of RNA-RNA mediated regulation in prokaryotes (Waters & Storz 2009; Storz et al. 2011; Updegrove et al. 2015), no study has yet evaluated the significance of unfavorable interactions between non-regulatory RNAs. We have tested the crosstalk effects *in silico* on available bacterial and archaeal genomes by extracting core mRNAs and core ncRNA genes. We also collected available prokaryotic gene expression data to detect the crosstalk influence on protein expression. Furthermore, we designed 13 green fluorescence

protein (GFP) reporter mRNAs and inserted these into *E. coli* to test the influence of avoidance on GFP expression.

Our results show that crosstalk avoidance is a widespread phenomenon in bacteria and archaea, and is supported by an evolutionarily conserved avoidance signal. We call this property RNA avoidance (or extrinsic RNA avoidance) which proves a reduced capacity for core mRNAs/ncRNAs binding (Chapter IV). We detected this reduced capacity by investigating binding MFE of mRNA-ncRNA pairs. Furthermore, our GFP reporter assay results show that we can accurately control gene expression levels by accounting for mRNA-ncRNA avoidance. Our model complements the current bacterial protein expression models by explaining the variance among protein-per-mRNA ratio that is not accounted for by the other two general factors: mRNA secondary structure and codon bias.

However, we found some exceptions during the evolutionary conservation analyses (Chapter IV). For example, bacterial phylum like *Planctomycetes* and *Cyanobacteria* have lower avoidance conservation than the others, 0% and ~45% respectively. We explained this discrepancy using an 'intrinsic avoidance' model, which invokes G+C difference of mRNAs and ncRNAs as a justification for lack of 'extrinsic' avoidance (Chapter IV). A similar mechanism with our 'intrinsic avoidance', the politeness hypothesis, was proposed which explains purine loading in thermophilic bacteria limits distracting RNA-RNA interactions (Lao & Forsdyke 2000). It states mRNAs of thermophilic bacteria should contain more purine to decrease the rate of RNA-RNA interactions (Lao & Forsdyke 2000). However, the politeness hypothesis does not explicitly describe the type of avoided RNA-RNA interactions or predict the influence of such interactions on protein expression. We suggested that an avoidance of RNA interactions is explicitly among ncRNAs and mRNAs, which was not grasped by the politeness.

Besides extrinsic (i.e. binding MFE signals) and intrinsic (i.e. G+C signals) features of RNAs, there may be some other biological reasons for lack of avoidance. There are a few assumptions that we considered in RNA avoidance model, including the co-occurrence and abundance of

interaction partners. In prokaryotes, translation and transcription occur in same place without any separation by internal membranes (Gowrishankar & Harinarayanan 2004), and even nascent mRNAs are translated during transcription (Laursen et al. 2005). Thus, all core ncRNAs (e.g. rRNAs and tRNAs) and mRNAs seem to exist in the same place (i.e. a crowded cytosol) which is also a justification for proximity of interaction partners, essential for successful binding (Waters & Storz 2009). Both core ncRNAs and mRNAs are quite abundant in prokaryotic cellular environment (Chapter II and IV). *Planctomycetes* members contain a membrane bounded nucleus (Lindsay et al. 2001; Fuerst 2005; Gottshall et al. 2014; Fuerst & Sagulenko 2011). Translation and transcription occur at different cell compartments in *Planctomycetes* strains *(Gottshall et al. 2014).* This defies the first assumption for RNA avoidance which could be an explanation for lack of avoidance in *Planctomycetes,* where the separation of abundant mRNAs from abundant ncRNAs serves as a 'spatial avoidance' mechanism that prevents crosstalk RNA-RNA interactions.

The *Cyanobacteria* results may also be partially explained by some biological insights. The members of this phylum are photosynthetic, and some of them can also fix atmospheric nitrogen in specialised cells called heterocysts, which is a strategy to protect nitrogenase from oxygen (Stanier & Cohen-Bazire 1977; Adams & Duggan 1999). Besides heterocysts, they also have 'akinete cells' that are specialised climate-resistant spores (Adams & Duggan 1999). Therefore, the RNA avoidance may not be necessary for those RNAs which are not co-transcribed, because of the differentiation of *Cyanobacteria* cells. A similar observation was performed in mammals which shows native regulatory miRNAs have evolved to selectively avoid matching sites on mRNAs if they are co-transcribed in differentiated cells (Farh et al. 2005). In short, the cellular specialization of *Cyanobacterial* cells may reduce the necessity of avoidance and serves as a 'temporal avoidance' mechanism.

In conclusion, our results show that the RNA avoidance is one of the major factors that influence protein expression in prokaryotes. The regulatory effects of ncRNAs on protein expression are already well-defined phenomena, but for the first time in RNA literature we propose that

mRNAs produce more protein if they do not readily bind (i.e. produce less stable interactions) to abundant non-regulatory core ncRNAs. We also prove that this can be detected as an avoidance signal (i.e. as a reduced capacity for core mRNAs/ncRNAs binding) in prokaryotic genomes. This shows an evolutionary selection for a high avoidance mRNAs like codon selection and low secondary structure of mRNAs.

5.2 Future perspectives

In the following sections, we present some future directions for another transcriptomics study to directly apply our phylogeny-informed sampling. We also summarize some possible downstream analyses related to the RNA avoidance.

5.2.1 Generating DNA-seq and RNA-seq data for AREBA-II

As we mentioned, the current RNA-seq data is heavily biased towards model organisms (or pathogens). In AREBA-II, we will sample from some underrepresented clades of life using phylogeny-informed sampling.

We have started by generating DNA-seq and RNA-seq data from *Planctomycetes* (a bacterial phylum) and *Halococcus* (an archaeal genus) members (Figure 5.2). *Planctomycetes* occur in soils, marine and freshwater aquatic habitats, and in some extreme habitats (Schlesner 1994). They reproduce by budding (Fuerst & Webb 1991; Devos & Reynaud 2010; Franzmann & Skerman 1984) and are known for their eukaryote-like features (e.g. internal membranes) (Franzmann & Skerman 1984; Fuerst 2005). *Planctomycetes* have cellular compartmentalization, which means that their cells are divided into different compartments by membranes (Lindsay et al. 2001; Fuerst 2005; Gottshall et al. 2014). This also causes segregation of transcription and translation events which are normally concurrent in bacteria (Gottshall et al. 2014). Furthermore, all *Planctomycetes* strains are missing peptidoglycan, an almost universal polymer found in bacterial cell wall (Lindsay et al. 2001; Fuerst 2005; Fuerst 1995). In summary, *Planctomycetes* cell structure differs from that of all other known prokaryotes because of the existence of internal membranes and peptidoglycan-less cell wall.

We have selected five strains of the *Planctomycetes* phylum (*G. obscuriglobus, Gemmata-like str. JW3-8s0, CJuql4, JW11-2f5* and *JW9-3f1*), which are inside the optimal phylogenetic distance (i.e. the Goldilocks zone) (Figure 5.1) and have some experimental advantages (e.g. easy to acquire and grow). Another important feature of the selected strains are their genome sizes which are quite large for bacterial strains (>= 9 Mb). Recent studies on small genomes also reported a huge repository of ncRNA genes (Chen et al. 2016; Güell et al. 2009), especially antisense regulatory ncRNAs. However, their functions are disputed and may be a result of high A+T content in their genomes (Lloréns-Rico et al. 2016). Yet, large genome sizes may reveal different ncRNA repositories. Therefore, phylogeny-informed selection ensures that these signals are correctly amplified for a comparative analysis.

Although some of the *Planctomycetes* strains already have draft genome sequences, we have created high quality DNA-seq data using Illumina and PacBio platforms with future RNA-seq experiments planned. We have currently assembled *de-novo* genomes for three strains. A recently published draft genome of *Gemmata massiliana* (*G. massiliana*) (Aghnatios et al. 2015) is also inside the optimal phylogenetic zone, so it can support our current sampling with another high quality draft genome (9.2 Mb consisting of 22 scaffolds), but this was not selected for our RNA-seq experiment as it was not available at the time.

Halococcus strains are members of *Euryarchaeota* and naturally live in high salt levels (Oren et al. 2009). All of the six selected *Halococcus* strains have draft genomes, and are located in the Goldilocks zone. We will generate both DNA-seq (for five strains) and RNA-seq data for all six strains. Unfortunately, as the available draft genomes are low quality and highly fragmented, we decided it would be better to re-sequence to acquire finished genomes, since the number of recovered genes from draft genomes are usually lower than the actual number of genes (Gurevich et al. 2013).

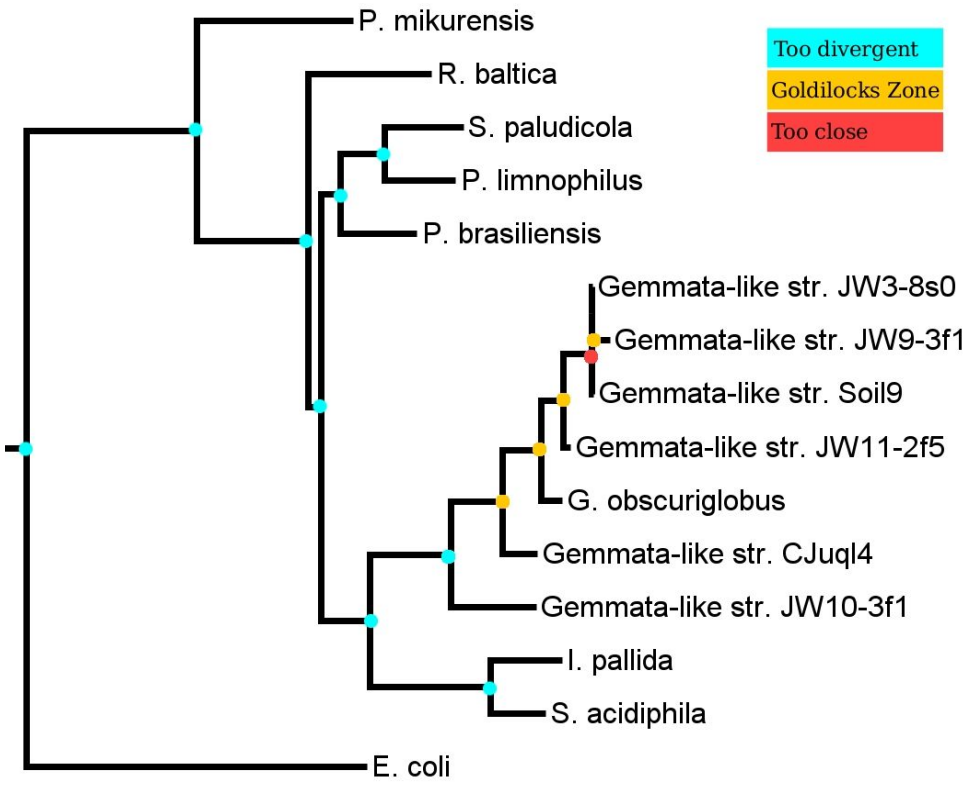

Figure 5.1. A maximum likelihood tree, generated by Phyml (Guindon et al. 2009), of the selected *Planctomycetes* strains is seen (*E. coli* as an outgroup). We aligned the ribosomal small subunit (SSU) RNAs of these strains using *cmalign* program *(Nawrocki et al. 2009)*. The yellow marked nodes show the Goldilocks zone strains including the selected strains for AREBA-II. The light-blue marked nodes show the connections of divergent strains while a single red dot shows the strains that are too similar for a comparative analysis.

In conclusion, sampling two different clades from bacteria and archaea will be a direct application of phylogeny-informed approach. We believe this will further illuminate genomic dark matter in large-genome-size-bacteria and underrepresented archaea. Generating RNA-seq data for both *Planctomycetes* and *Halococcus* strains will extend our knowledge on ncRNA biology (Figure 5.2).

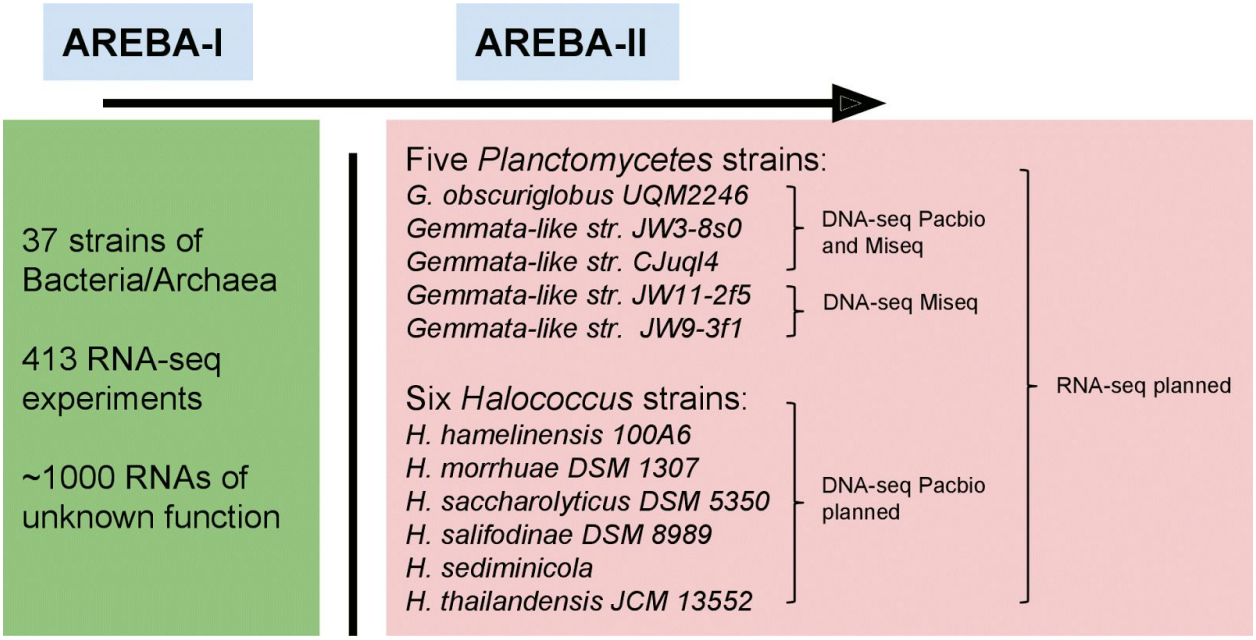

Figure 5.2 The AREBA project phase one and two overview**.**

5.2.2 The RNA avoidance hypothesis reloaded

It seems RNA avoidance is a widespread phenomenon in prokaryotes. Therefore, it is worth doing further work on this subject. As we mentioned that we detected a lack of (extrinsic) avoidance when compared to other bacterial strains (Chapter IV) (Umu et al. 2015). On the other hand, there are not many sequenced *Planctomycetes* genomes available to totally reject RNA avoidance for *Planctomycetes* strains (there are only 4 strains available). Therefore, it would be logical to deeply investigate RNA avoidance on newly sequenced *Planctomycetes* strains, which we plan to create DNA-seq and RNA-seq data for AREBA-II.

In contrast, we may observe RNA avoidance in some eukaryotes (especially in lower eukaryotes). For example, abundant regulatory RNAs (e.g. eukaryotic small RNAs) and abundant lncRNAs may avoid interaction with mRNAs. There is also novel class of small RNAs, called tRNA-derived RNA fragments (tRFs), available in eukaryotes (Lee et al. 2009; Goodarzi et al. 2015) which proves regulatory interactions between mRNAs and tRNAs are possible. Furthermore, an analogous miRNA-mRNA avoidance was observed in mammals (Farh et al.

2005). Therefore, it would be logical to do further work in eukaryotes. Eukaryotic mitochondria and chloroplast are also ideal candidates for an investigation due to their resemblance to prokaryotic cells.

Besides these, avoidance may have implications for G+C bias and sRNA evolution in prokaryotes. For example, G+C distributions vary among prokaryotic genomes (Hildebrand et al. 2010) as well as between ncRNA genes and mRNAs, which cannot be explained by adaptation for optimal growth temperature (Hurst & Merchant 2001; Zeldovich et al. 2007). The 'politeness' hypothesis states that purine-loading restricts distracting RNA-RNA interactions in thermophilic bacteria (Lao & Forsdyke 2000). Therefore, (intrinsic) RNA avoidance may be a factor that constrains G+C distribution in genomes and optimizes G+C variation to decrease crosstalk RNA interactions. In fact, we proposed the intrinsic avoidance features of mRNAs (Chapter IV) based on this idea. Moreover, sRNA evolution in bacteria may be affected by RNA avoidance. It is believed that pervasive transcription products are subject to evolutionary selection in order to create new genes (Wade & Grainger 2014). Thus, RNA avoidance can be one of the factors that drives this selection.

It seems RNA interaction avoidance is an important factor to optimize transformed genes, so it can be used as a complementary factor to design genes for synthetic biology. Therefore, an algorithm that can produce the most optimal codon configuration by including all major factors to increase protein output for various host cells with different host ncRNA gene sets will be useful. Such an algorithm might also be useful to fine-tune final protein counts rather than increasing protein outputs. The current avoidance model also assumes the equal contribution of core ncRNAs to the final avoidance score and the region of avoidance is limited to 5` ends of mRNAs (Chapter IV), which can be fine-tuned for a novel gene design algorithm.

In conclusion, RNA avoidance may have applications on various other studies in the future, from answering evolutionary biology questions, to designing new bioinformatics tools (e.g. gene design tools). Furthermore, the field of synthetic biology is emerging to solve some fundamental

problems of the world such as producing biofuels, creating drugs and tissues for medical purposes (Khalil & Collins 2010; Purnick & Weiss 2009). RNAs play a central role in synthetic biology not only for carrying protein coding information but also for their versatility and regulatory properties (Chappell et al. 2013). Thus, it seems RNA avoidance can be an important addition to the discipline of engineering gene expression inside cells.

REFERENCES

Adams, D.G. & Duggan, P.S., 1999. Tansley Review No. 107. Heterocyst and akinete differentiation in cyanobacteria. *The New phytologist*, 144(1), pp.3–33.

Aghnatios, R. et al., 2015. Draft genome of Gemmata massiliana sp. nov, a water-borne Planctomycetes species exhibiting two variants. *Standards in genomic sciences*, 10, p.120.

Akashi, H., 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, 136(3), pp.927–935.

Andronescu, M. et al., 2003. RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic acids research*, 31(13), pp.3416–3422.

van Bakel, H. et al., 2010. Most "Dark Matter" Transcripts Are Associated With Known Genes. *PLoS biology*, 8(5), p.e1000371.

Barquist, L. & Vogel, J., 2015. Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annual review of genetics*, 49, pp.367–394.

Busch, A., Richter, A.S. & Backofen, R., 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* , 24(24), pp.2849–2856.

Chamary, J.V. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9), p.R75.

Chappell, J. et al., 2013. The centrality of RNA for engineering gene expression. *Biotechnology journal*, 8(12), pp.1379–1395.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Chou, C.-H. et al., 2016. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research*, 44(D1), pp.D239–47.

Clark, M.B. et al., 2011. The reality of pervasive transcription. *PLoS biology*, 9(7), p.e1000625; discussion e1001102.

Cohen, O. et al., 2016. Comparative transcriptomics across the prokaryotic tree of life. *Nucleic acids research*. Available at: http://dx.doi.org/10.1093/nar/gkw394.

Conesa, A. & Götz, S., 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, 2008, p.619832.

Croucher, N.J. & Thomson, N.R., 2010. Studying bacterial transcriptomes using RNA-seq. *Current opinion in microbiology*, 13(5), pp.619–624.

Devos, D.P. & Reynaud, E.G., 2010. Evolution. Intermediate steps. *Science*, 330(6008), pp.1187–1188.

Dieterich, C. & Stadler, P.F., 2012. Computational biology of RNA interactions. *Wiley interdisciplinary reviews. RNA*, 4(1), pp.107–120.

Farh, K.K.-H. et al., 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755), pp.1817–1821.

Franzmann, P.D. & Skerman, V.B., 1984. Gemmata obscuriglobus, a new genus and species of the budding bacteria. *Antonie van Leeuwenhoek*, 50(3), pp.261–268.

Fuerst, J.A., 2005. Intracellular compartmentation in planctomycetes. *Annual review of microbiology*, 59, pp.299–328.

Fuerst, J.A., 1995. The planctomycetes: emerging models for microbial ecology, evolution and cell biology. *Microbiology*, 141 (Pt 7), pp.1493–1506.

Fuerst, J.A. & Sagulenko, E., 2011. Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nature reviews. Microbiology*, 9(6), pp.403–413.

Fuerst, J.A. & Webb, R.I., 1991. Membrane-bounded nucleoid in the eubacterium Gemmatata obscuriglobus. *Proceedings of the National Academy of Sciences*, 88(18), pp.8184–8188.

Gardner, P.P. & Eldai, H., 2015. Annotating RNA motifs in sequences and alignments. *Nucleic acids research*, 43(2), pp.691–698.

Gardner, P.P. & Giegerich, R., 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC bioinformatics*, 5(1), p.18.

Gaspar, P. et al., 2013. mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic acids research*, 41(6), pp.e73–e73.

Goodarzi, H. et al., 2015. Endogenous tRNA-Derived Fragments Suppress Breast Cancer

Progression via YBX1 Displacement. *Cell*, 161(4), pp.790–802.

Goodman, D.B., Church, G.M. & Kosuri, S., 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(6157), pp.475–479.

Gottshall, E.Y. et al., 2014. Spatially segregated transcription and translation in cells of the endomembrane-containing bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30), pp.11067–11072.

Gowrishankar, J. & Harinarayanan, R., 2004. Why is transcription coupled to translation in bacteria? *Molecular microbiology*, 54(3), pp.598–603.

Gruber, A.R. et al., 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic acids research*, 35(Web Server issue), pp.W335–8.

Güell, M. et al., 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature reviews. Microbiology*, 9(9), pp.658–669.

Güell, M. et al., 2009. Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957), pp.1268–1271.

Guindon, S. et al., 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology* , 537, pp.113–137.

Gurevich, A. et al., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* , 29(8), pp.1072–1075.

Gu, W. et al., 2014. The impact of RNA structure on coding sequence evolution in both bacteria and eukaryotes. *BMC evolutionary biology*, 14, p.87.

Hildebrand, F., Meyer, A. & Eyre-Walker, A., 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS genetics*, 6(9), p.e1001107.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1), pp.44–57.

Hurst, L.D. & Merchant, A.R., 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings. Biological sciences / The Royal Society*, 268(1466), pp.493–497.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), pp.13–34.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology*,

151(3), pp.389–409.

Kapranov, P. et al., 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830), pp.1484–1488.

Kato, Y. et al., 2010. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* , 26(18), pp.i460–6.

Khalil, A.S. & Collins, J.J., 2010. Synthetic biology: applications come of age. *Nature reviews. Genetics*, 11(5), pp.367–379.

Krüger, J. & Rehmsmeier, M., 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research*, 34(Web Server issue), pp.W451–4.

Kudla, G. et al., 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* , 324(5924), pp.255–258.

Kwon, T. et al., 2014. Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. *Journal of proteome research*, 13(5), pp.2370–2380.

Lai, D. & Meyer, I.M., 2015. A comprehensive comparison of general RNA-RNA interaction prediction methods. *Nucleic acids research*. Available at: http://dx.doi.org/10.1093/nar/gkv1477.

Lao, P.J. & Forsdyke, D.R., 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome research*, 10(2), pp.228–236.

Laursen, B.S. et al., 2005. Initiation of protein synthesis in bacteria. *Microbiology and molecular biology reviews: MMBR*, 69(1), pp.101–123.

Layton, D.M. & Bundschuh, R., 2005. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic acids research*, 33(2), pp.519–524.

Lee, Y.S. et al., 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & development*, 23(22), pp.2639–2649.

Lindgreen, S. et al., 2014. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS computational biology*, 10(10), p.e1003907.

Lindsay, M.R. et al., 2001. Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Archives of microbiology*, 175(6), pp.413–429.

Lloréns-Rico, V. et al., 2016. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science advances*, 2(3), p.e1501363.

Lorenz, R. et al., 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology: AMB*, 6,

p.26.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), p.550.

Lu, P. et al., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25(1), pp.117–124.

Maier, T. et al., 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular systems biology*, 7, p.511.

Mathews, D.H. et al., 1999. Expanded sequence dependence of. *A Polymorphism Increases GR-A*.

McCarthy, D.J., Chen, Y. & Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10), pp.4288–4297.

Mistry, J. et al., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41(12), p.e121.

Nawrocki, E.P. & Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , 29(22), pp.2933–2935.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R., 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* , 25(10), pp.1335–1337.

Oren, A., Arahal, D.R. & Ventosa, A., 2009. Emended descriptions of genera of the family Halobacteriaceae. *International journal of systematic and evolutionary microbiology*, 59(Pt 3), pp.637–642.

Pain, A. et al., 2015. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5), pp.509–513.

Palazzo, A.F. & Gregory, T.R., 2014. The case for junk DNA. *PLoS genetics*, 10(5), p.e1004351.

Pauli, A., Valen, E. & Schier, A.F., 2015. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 37(1), pp.103–112.

Pelletier, J. & Sonenberg, N., 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 65(6), pp.576–581.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1), pp.32–42.

Poolsap, U. et al., 2011. Using binding profiles to predict binding sites of target RNAs. *Journal of bioinformatics and computational biology*, 9(6), pp.697–713.

Purnick, P.E.M. & Weiss, R., 2009. The second wave of synthetic biology: from modules to systems. *Nature reviews. Molecular cell biology*, 10(6), pp.410–422.

Richter, A.S. & Backofen, R., 2012. Accessibility and conservation: General features of bacterial small RNA-mRNA interactions? *RNA biology*, 9(7), pp.954–965.

Schlesner, H., 1994. The Development of Media Suitable for the Microorganisms Morphologically Resembling Planctomyces spp., Pirellula spp., and other Planctomycetales from Various Aquatic Habitats Using Dilute Media. *Systematic and applied microbiology*, 17(1), pp.135–145.

Sharma, C.M. et al., 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286), pp.250–255.

de Sousa Abreu, R. et al., 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12), pp.1512–1526.

Stanier, R.Y. & Cohen-Bazire, G., 1977. Phototrophic prokaryotes: the cyanobacteria. *Annual review of microbiology*, 31, pp.225–274.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Tafer, H. & Hofacker, I.L., 2008. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* , 24(22), pp.2657–2663.

Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.

Tuller, T. et al., 2010. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8), pp.3645–3650.

Umu, S.U. et al., 2015. *Natural avoidance of stochastic mRNA:ncRNA interactions can be harnessed to control protein expression levels*, Available at: http://dx.doi.org/10.1101/033613.

Updegrove, T.B., Shabalina, S.A. & Storz, G., 2015. How do base-pairing small RNAs evolve? *FEMS microbiology reviews*, 39(3), pp.379–391.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Wade, J.T. & Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of

bacterial transcriptomes. *Nature reviews. Microbiology*, 12(9), pp.647–653.

Wang, J. et al., 2016. sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic acids research*, 44(D1), pp.D248–53.

Wang, K.C. et al., 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341), pp.120–124.

Washietl, S. et al., 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* , 17(4), pp.578–594.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

Wheeler, T.J. & Eddy, S.R., 2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* , 29(19), pp.2487–2489.

Workman, C. & Krogh, A., 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24), pp.4816–4822.

Zeldovich, K.B., Berezovsky, I.N. & Shakhnovich, E.I., 2007. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS computational biology*, 3(1), p.e5.